MODEL SELECTION IN LOGISTIC REGRESSION

MARIUS KWEMOU[(1)], MARIE-LUCE TAUPIN[(1)(2)], ANNE-SOPHIE TOCQUET[(1)]

Abstract. This paper is devoted to model selection in logistic regression. We extend the model selection principle introduced by Birgé and Massart (2001) to logistic regression model. This selection is done by using penalized maximum likelihood criteria. We propose in this context a completely data-driven criteria based on the slope heuristics. We prove non asymptotic oracle inequalities for selected estimators. Theoretical results are illustrated through simulation studies.

Keywords: logistic regression, model selection, projection.
AMS 2000 MSC: Primary 62J02, 62F12, Secondary 62G05, 62G20.

1. Introduction

Consider the following generalization of the logistic regression model : let $(Y_1, x_1), \cdots, (Y_n, x_n)$, be a sample of size n such that $(Y_i, x_i) \in \{0, 1\} \times \mathcal{X}$ and

$$\mathbb{E}_{f_0}(Y_i) = \pi_{f_0}(x_i) = \frac{\exp f_0(x_i)}{1 + \exp f_0(x_i)},$$

where f_0 is an unknown function to be estimated and the design points $x_1, ..., x_n$ are deterministic. This model can be viewed as a nonparametric version of the "classical" logistic model which relies on the assumption that $x_i \in \mathbb{R}^d$, and that there exists $\beta_0 \in \mathbb{R}^d$ such that $f_0(x_i) = \beta_0^\top x_i$.

Logistic regression is a widely used model for predicting the outcome of binary dependent variable. For example logistic model can be used in medical study to predict the probability that a patient has a given disease (e.g. cancer), using observed characteristics (explanatory variables) of the patient such as weight, age, patient's gender *etc*. However in the presence of numerous explanatory variables with potential influence, one would like to use only a few number of variables, for the sake of interpretability or to avoid overfitting. But it is not always obvious to choose the adequate variables. This is the well-known problem of variables selection or model selection.

In this paper, the unknown function f_0 is not specified and not necessarily linear. Our aim is to estimate f_0 by a linear combination of given functions, often called dictionary. The dictionary can be a basis of functions, for instance spline or polynomial basis.

A nonparametric version of the classical logistic model has already been considered by Hastie (1983), where a nonparametric estimator of f_0 is proposed using local maximum likelihood. The problem of nonparametric estimation in additive regression model is well known and deeply studied. But in logistic regression model it is less studied. One can cite for instance Lu (2006), Vexler (2006), Fan *et al.* (1998), Farmen (1996), Raghavan (1993), and Cox (1990).

Recently few papers deal with model selection or nonparametric estimation in logistic regression using ℓ_1 penalized contrast Bunea (2008), Bach (2010), van de Geer (2008), Kwemou (2012). Among them, some establish non asymptotic oracle inequalities that hold even in high dimensional setting. When the dimension of $\mathcal{X}$ is high, that is greater than dozen, such ℓ_1 penalized contrast estimators are known to provide reasonably good results. When the dimension of $\mathcal{X}$ is small, it is often better to choose different penalty functions. One classical penalty function is what we call ℓ_0 penalization. Such penalty functions, built as increasing function of the dimension of $\mathcal{X}$, usually refers to model selection. The last decades have witnessed a growing interest in the model selection problem since the seminal works of Akaike (1973), Schwarz (1978). In additive regression one can cite among the others Baraud (2000a), Birgé and Massart (2001), Yang (1999), in density estimation Birgé (2014), Castellan (2003a) and in segmentation problem Lebarbier (2005), Durot *et al.* (2009), and Braun *et al.* (2000). All the previously cited papers use ℓ_0 penalized contrast to perform model selection. But model selection procedures based on penalized maximum likelihood estimators in logistic regression are less studied in the literature.

In this paper we focus on model selection using ℓ_0 penalized contrast for logistic regression model and in this context we state non asymptotic oracle inequalities. More precisely, given some collection functions, we consider estimators of f_0 built as linear combination of the functions. The point that the true function is not supposed to be linear combination of those functions, but we expect that the spaces of linear combination of those functions would provide suitable approximation spaces. Thus, to this collection of functions, we associate a collection of estimators of f_0. Our aim is to propose a data driven procedure, based on penalized criterion, which will be able to choose the "best" estimator among the collection of estimators, using ℓ_0 penalty functions.

The collection of estimators is built using minimisation of the opposite of logarithm likelihood. The properties of estimators are described in term of Kullback-Leibler divergence and the empirical L_2 norm. Our results can be splitted into two parts.

First, in a general model selection framework, with general collection of functions we provide a completely data driven procedure that automatically selects the best model among the collection. We state non asymptotic oracle inequalities for Kullback-Leibler divergence and the empirical L_2 norm between the selected estimator and the true function f_0. The estimation procedure relies on the building of a suitable penalty function, suitable in the sense that it performs best risks and suitable in the sense that it does not depend on the unknown smoothness parameters of the true function f_0. But, the penalty function depends on a bound related to target function f_0. This can be seen as the price to pay for the generality. It comes from needed links between Kullback-Leibler divergence and empirical L_2 norm.

Second, we consider the specific case of collection of piecewise functions which provide estimator of type regressogram. In this case, we exhibit a completely data driven penalty, free from f_0. The model selection procedure based on this penalty provides an adaptive estimator and state a non asymptotic oracle inequality for Hellinger distance and the empirical L_2 norm

between the selected estimator and the true function f_0. In the case of piecewise constant functions basis, the connection between Kullback-Leibler divergence and the empirical L_2 norm are obtained without bound on the true function f_0. This last result is of great interest for example in segmentation study, where the target function is piecewise constant or can be well approximated by piecewise constant functions.

Those theoretical results are illustrated through simulation studies. In particular we show that our model selection procedure (with the suitable penalty) have good non asymptotic properties as compared to usual known criteria such as AIC and BIC. A great attention has been made on the practical calibration of the penalty function. This practical calibration is mainly based on the ideas of what is usually referred as slope heuristic as proposed in Birgé and Massart (2007) and developed in Arlot and Massart (2009).

The paper is organized as follow. In Section 2 we set our framework and describe our estimation procedure. In Section 3 we define the model selection procedure and state the oracle inequalities in the general framework. Section 4 is devoted to regressogram selection, in this section, we establish a bound of the Hellinger risk between the selected model and the target function. The simulation study is reported in Section 5. The proofs of the results are postponed to Section 6 and 7.

2. Model and framework

Let $(Y_1, x_1), \cdots, (Y_n, x_n)$, be a sample of size n such that $(Y_i, x_i) \in \{0,1\} \times \mathcal{X}$. Throughout the paper, we consider a fixed design setting *i.e.* $x_1, \ldots, x_n$ are considered as deterministic. In this setting, consider the extension of the "classical" logistic regression model (2.1) where we aim at estimating the unknown function f_0 in

$$\mathbb{E}_{f_0}(Y_i) = \pi_{f_0}(x_i) = \frac{\exp f_0(x_i)}{1 + \exp f_0(x_i)}. \tag{2.1}$$

We propose to estimate the unknown function f_0 by model selection. This model selection is performed using penalized maximum likelihood estimators. In the following we denote by $\mathbb{P}_{f_0}(x_1)$ the distribution of Y_1 and by $\mathbb{P}_{f_0}^{(n)}(x_1, \cdots, x_n)$ the distribution of $(Y_1, \ldots, Y_n)$ under Model (2.1). Since the variables Y_i's are independent random variables,

$$\mathbb{P}_{f_0}^{(n)}(x_1, \cdots, x_n) = \Pi_{i=1}^n \mathbb{P}_{f_0}(x_i) = \prod_{i=1}^n \pi_{f_0}(x_i)^{Y_i}(1 - \pi_{f_0}(x_i))^{1-Y_i}.$$

It follows that for a function f mapping $\mathcal{X}$ into $\mathbb{R}$, the likelihood is defined as:

$$L_n(f) = \mathbb{P}_f^{(n)}(x_1, \cdots, x_n) = \prod_{i=1}^n \pi_f(x_i)^{Y_i}(1 - \pi_f(x_i))^{1-Y_i},$$

where

$$\pi_f(x_i) = \frac{\exp(f(x_i))}{1 + \exp(f(x_i))}. \tag{2.2}$$

We choose the opposite of the log-likelihood as the estimation criterion that is

$$\gamma_n(f) = -\frac{1}{n}\log(L_n(f)) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\log(1+e^{f(x_i)}) - Y_i f(x_i)\Big\}. \tag{2.3}$$

Associated to this estimation criterion we consider the Kullback-Leibler information divergence $\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f})$ defined as

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f}) = \frac{1}{n}\int \log\left(\frac{\mathbb{P}^{(n)}_{f_0}}{\mathbb{P}^{(n)}_{f}}\right) d\mathbb{P}^{(n)}_{f_0}.$$

The loss function is the excess risk, defined as

$$\mathcal{E}(f) := \gamma(f) - \gamma(f_0) \text{ where, for any } f, \quad \gamma(f) = \mathbb{E}_{f_0}[\gamma_n(f)]. \tag{2.4}$$

Easy calculations show that the excess risk is linked to the Kullback-Leibler information divergence through the relation

$$\mathcal{E}(f) = \gamma(f) - \gamma(f_0) = \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f}).$$

It follows that, f_0 minimizes the excess risk, that is

$$f_0 = \arg\min_f \gamma(f).$$

As usual, one can not estimate f_0 by the minimizer of $\gamma_n(f)$ over any functions space, since it is infinite. The usual way is to minimize $\gamma_n(f)$ over a finite dimensional collections of models, associated to a finite dictionary of functions $\phi_j : \mathcal{X} \to \mathbb{R}$

$$\mathcal{D} = \{\phi_1, \ldots, \phi_M\}.$$

For the sake of simplicity we will suppose that $\mathcal{D}$ is a orthonormal basis of functions. Indeed, if $\mathcal{D}$ is not an orthonormal basis of functions, we can always find an orthonormal basis of functions $\mathcal{D}' = \{\psi_1, \ldots, \psi_{M'}\}$ such that

$$\langle \phi_1, \ldots, \phi_M \rangle = \langle \psi_1, \ldots, \psi_{M'} \rangle.$$

Let $\mathcal{M}$ the set of all subsets $m \subset \{1, \ldots, M\}$. For every $m \in \mathcal{M}$, we call $\mathcal{S}_m$ the model

$$\mathcal{S}_m := \Big\{ f_\beta = \sum_{j \in m} \beta_j \phi_j \Big\} \tag{2.5}$$

and D_m the dimension of the span of $\{\phi_j, j \in m\}$. Given the countable collection of models $\{\mathcal{S}_m\}_{m\in\mathcal{M}}$, we define $\{\hat{f}_m\}_{m\in\mathcal{M}}$ the corresponding estimators, *i.e.* the estimators obtaining by minimizing γ_n over each model $\mathcal{S}_m$. For each $m \in \mathcal{M}$, $\hat{f}_m$ is defined by

$$\hat{f}_m = \arg\min_{t \in \mathcal{S}_m} \gamma_n(t). \tag{2.6}$$

Our aim is choose the "best" estimator among this collection of estimators, in the sense that it minimizes the risk. In many cases, it is not easy to choose the "best" model. Indeed, a model with small dimension tends to be efficient from estimation point of view whereas it could

be far from the "true" model. On the other side, a more complex model easily fits data but the estimates have poor predictive performance (overfitting). We thus expect that this best estimator mimics what is usually called the oracle defined as

$$m^* = \arg\min_{m\in\mathcal{M}} \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m}). \tag{2.7}$$

Unfortunately, both, minimizing the risk and minimazing the kulback-leibler divergence, require the knowledge of the true (unknown) function f_0 to be estimated.

Our goal is to develop a data driven strategy based on data, that automatically selects the best estimator among the collection, this best estimator having a risk as close as possible to the oracle risk, that is the risk of $\hat{f}_{m^*}$. In this context, our strategy follows the lines of model selection as developed by Birgé and Massart (2001). We also refer to the book Massart (2007) for further details on model selection.

We use penalized maximum likelihood estimator for choosing some data-dependent $\hat{m}$ nearly as good as the ideal choice m^*. More precisely, the idea is to select $\hat{m}$ as a minimizer of the penalized criterion

$$\hat{m} = \arg\min_{m\in\mathcal{M}} \left\{\gamma_n(\hat{f}_m) + \mathrm{pen}(m)\right\}, \tag{2.8}$$

where pen : $\mathcal{M} \longrightarrow \mathbb{R}^+$ is a data driven penalty function. The estimation properties of $\hat{f}_m$ are evaluated by non asymptotic bounds of a risk associated to a suitable chosen loss function. The great challenge is choosing the penalty function such that the selected model $\hat{m}$ is nearly as good as the oracle m^*. This penalty term is classically based on the idea that

$$m^* = \arg\min_{m\in\mathcal{M}} \mathbb{E}_{f_0}\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m}) = \arg\min_{m\in\mathcal{M}} \left[\mathbb{E}_{f_0}\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathbb{E}_{f_0}\mathcal{K}(\mathbb{P}^{(n)}_{f_m}, \mathbb{P}^{(n)}_{\hat{f}_m})\right]$$

where f_m is defined as

$$f_m = \arg\min_{t\in S_m} \gamma(t).$$

Our goal is to build a penalty function such that the selected model $\hat{m}$ fulfills an oracle inequality:

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leq C_n \inf_{m\in\mathcal{M}} \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m}) + R_n.$$

This inequality is expected to hold either in expectation or with high probability, where C_n is as close to 1 as possible and R_n is a remainder term negligible compared to $\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{m*}})$.

In the following we consider two separated case. First we consider general collection of models under boundedness assumption. Second we consider the specific case of regressogram collection.

3. Oracle inequality for general models collection under boundedness assumption

Consider model (2.1) and $(\mathcal{S}_m)_{m\in\mathcal{M}}$ a collection of models defined by (2.5). Let $C_0 > 0$ and $\mathbb{L}_\infty(C_0) = \left\{f : \mathcal{X} \to \mathbb{R},\ \max_{1\leqslant i\leqslant n} |f(x_i)| \leqslant C_0\right\}$. For $m \in \mathcal{M}$, γ_n given in (2.3), and γ is given by

(2.4), we define

$$\hat{f}_m = \arg \min_{t \in \mathcal{S}_m \cap \mathbb{L}_\infty(C_0)} \gamma_n(t) \text{ and } f_m = \arg \min_{t \in \mathcal{S}_m \cap \mathbb{L}_\infty(C_0)} \gamma(t). \tag{3.9}$$

The first step consists in studying the estimation properties of $\hat{f}_m$ for each m, as it is stated in the following proposition.

Proposition 3.1. *Let $C_0 > 0$ and $\mathcal{U}_0 = e^{C_0}/(1+e^{C_0})^2$. For $m \in \mathcal{M}$, let $\hat{f}_m$ and f_m as in (3.9). We have*

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m})] \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \frac{D_m}{2n\mathcal{U}_0^2}$$

This proposition says that the "best" estimator amoung the collection $\{\hat{f}_m\}_{m\in\mathcal{M}}$, in the sense of the Kullback-Leibler risk, is the one which makes a balance between the bias and the complexity of the model. In the ideal situation where f_0 belongs to $\mathcal{S}_m$, we have that

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m})] \leqslant \frac{1}{\mathcal{U}_0^2}\frac{D_m}{2n}.$$

To derive the model selection procedure we need the following assumption :

$(\mathbf{A_1})$ There exists a constant $0 < c_1 < \infty$ such that $\max_{1\leqslant i \leqslant n} |f_0(x_i)| \leqslant c_1$.

In the following theorem we propose a choice for the penalty function and we state non asymptotic risk bounds.

Theorem 3.1. *Given $C_0 > 0$, for $m \in \mathcal{M}$, let $\hat{f}_m$ and f_m be defined as (3.9). Let us denote $\| f \|_n^2 = n^{(-1)} \sum_{i=1}^n f^2(x_i)$. Let $\{L_m\}_{m\in\mathcal{M}}$ some positive numbers satisfying*

$$\Sigma = \sum_{m\in\mathcal{M}} \exp(-L_m D_m) < \infty.$$

We define pen : $\mathcal{M} \to \mathbb{R}_+$, such that, for $m \in \mathcal{M}$,

$$pen(m) \geqslant \lambda \frac{D_m}{n}\left(\frac{1}{2} + \sqrt{5L_m}\right)^2,$$

where λ is a positive constant depending on c_1. Under Assumption $(\mathbf{A_1})$ we have

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})] \leqslant C \inf_{m\in\mathcal{M}} \left\{\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + pen(m)\right\} + C_1\frac{\Sigma}{n}$$

and

$$\mathbb{E}_{f_0} \| \hat{f}_{\hat{m}} - f_0 \|_n^2 \leqslant C' \inf_{m\in\mathcal{M}} \left\{\| f_0 - f_m \|_n^2 + pen(m)\right\} + C_1'\frac{\Sigma}{n}.$$

where C, C', C_1, C_1' are constants depending on c_1 and C_0.

This theorem provides oracle inequalities for L_2-norm and for K-L divergence between the selected model and the true function. Provided that penalty has been properly chosen, one can bound the L_2-norm and the K-L divergence between the selected model and the true function. The inequalities in Theorem 3.1 are non-asymptotic inequalities in the sense that the result is obtain for a fixed n. This theorem is very general and does not make specific assumption on the dictionary. However, the penalty function depends on some unknown constant λ which depends on the bound of the true function f_0 through Condition (6.5). In practice this constant can be calibrated using "slope heuristics" proposed in Birgé and Massart (2007). In the following we will show how to obtain similar result with a penalty function not connected to the bound of the true unknown function f_0 in the regressogram case.

4. Regressogram functions

4.1. **Collection of models.** In this section we suppose (without loss of generality) that $f_0 : [0,1] \to \mathbb{R}$. For the sake of simplicity, we use the notation $f_0(x_i) = f_0(i)$ for every $i = 1, \dots, n$. Hence f_0 is defined from $\{1, \dots, n\}$ to $\mathbb{R}$. Let $\mathcal{M}$ be a collection of partitions of intervals of $\mathcal{X} = \{1, \dots, n\}$. For any $m \in \mathcal{M}$ and $J \in m$, let $\mathbb{I}_J$ denote the indicator function of J and S_m be the linear span of $\{\mathbb{I}_J, J \in m\}$. When all intervals have the same length, the partition is said regular, and is is irregular otherwise.

4.2. **Collection of estimators: regressogram.** For a fixed m, the minimizer $\hat{f}_m$ of the empirical contrast function γ_n, over S_m, is called the *regressogram*. That is, f_0 is estimated by $\hat{f}_m$ given by

$$\hat{f}_m = \arg\min_{f \in S_m} \gamma_n(f). \tag{4.10}$$

where γ_n is given by (2.3). Associated to S_m we have

$$f_m = \arg\min_{f \in S_m} \gamma(f) - \gamma(f_0) = \arg\min_{f \in S_m} \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f}). \tag{4.11}$$

In the specific case where S_m is the set of piecewise constant functions on some partition m, $\hat{f}_m$ and f_m are given by the following lemma.

Lemma 4.1. *For $m \in \mathcal{M}$, let f_m and $\hat{f}_m$ be defined by (4.11) and (4.10) respectively . Then, $f_m = \sum_{J \in m} \overline{f}^{(J)}_m \mathbb{I}_J$ and $\hat{f}_m = \sum_{J \in m} \hat{f}^{(J)}_m \mathbb{I}_J$ with*

$$\overline{f}^{(J)}_m = \log\left(\frac{\sum_{i \in J} \pi_{f_0}(x_i)}{|J|(1 - \sum_{i \in J} \pi_{f_0}(x_i)/|J|)}\right) \text{ and } \hat{f}^{(J)}_m = \log\left(\frac{\sum_{i \in J} Y_i}{|J|(1 - \sum_{i \in J} Y_i/|J|)}\right).$$

Moreover, $\pi_{f_m} = \sum_{J \in m} \pi^{(J)}_{f_m} \mathbb{I}_J$ and $\pi_{\hat{f}_m} = \sum_{J \in m} \pi^{(J)}_{\hat{f}_m} \mathbb{I}_J$ with

$$\pi^{(J)}_{f_m} = \frac{1}{|J|} \sum_{i \in J} \pi_{f_0}(x_i), \text{ and } \pi^{(J)}_{\hat{f}_m} = \frac{1}{|J|} \sum_{i \in J} Y_i.$$

Consequently, $\pi_{f_m} = \arg\min_{\pi \in S_m} \| \pi - \pi_{f_0} \|^2_n$ is the usual projection of π_{f_0} on to S_m.

4.3. **First bounds on $\hat{f}_m$.** Consider the following assumptions:

($\mathbf{A_2}$) There exists a constant $\rho > 0$ such that $\min_{i=1,\cdots,n} \pi_{f_0}(x_i) \geq \rho$ and $\min_{i=1,\cdots,n} [1 - \pi_{f_0}(x_i)] \geq \rho$.

Proposition 4.1. *Consider Model (2.1) and let $\hat{f}_m$ be defined by (4.10) with m such that for all $J \in m$, $|J| \geqslant \Gamma[\log(n)]^2$ for a positive constant Γ. Under Assumption ($\mathbf{A_2}$), for all $\delta > 0$ and $a > 1$, we have*

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m})] \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m})) + \frac{(1+\delta)D_m}{(1-\delta)^2 n} + \frac{\kappa(\Gamma, \rho, \delta)}{n^a}.$$

4.4. **Adaptive estimation and oracle inequality.** The following result provides an adaptive estimation of f_0 and a risk bound of the selected model.

Definition 4.1. *Let $\mathcal{M}$ be a collection of partitions of $\mathcal{X} = \{1, \ldots, n\}$ constructed on the partition m_f i.e. m_f is a refinement of every $m \in \mathcal{M}$.*

In other words, a partition m belongs to $\mathcal{M}$ if any element of m is the union of some elements of m_f. Thus S_{m_f} contains every model of the collection $\{S_m\}_{m\in\mathcal{M}}$.

Theorem 4.1. *Consider Model (2.1) under Assumption ($\mathbf{A_2}$). Let $\{S_m, m \in \mathcal{M}\}$ be a collection of models defined in Section 4.1 where $\mathcal{M}$ is a set of partitions constructed on the partition m_f such that*

$$\text{for all } J \in m_f, |J| \geq \Gamma \log^2(n), \tag{4.1}$$

where Γ is a positive constant. Let $(L_m)_{m\in\mathcal{M}}$ be some family of positive weights satisfying

$$\Sigma = \sum_{m\in\mathcal{M}} \exp(-L_m D_m) < +\infty. \tag{4.2}$$

Let $pen : \mathcal{M} \to \mathbb{R}_+$ satisfying for $m \in \mathcal{M}$, and for $\mu > 1$,

$$pen(m) \geqslant \mu \frac{D_m}{n}\left(1 + 6L_m + 8\sqrt{L_m}\right).$$

Let $\tilde{f} = \hat{f}_{\hat{m}}$ where

$$\hat{m} = \arg\min_{m\in\mathcal{M}} \left\{\gamma_n(\hat{f}_m) + pen(m)\right\},$$

then, for $C_\mu = 2\mu^{1/3}/(\mu^{1/3} - 1)$, we have

$$\mathbb{E}_{f_0}[h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\tilde{f}})] \leqslant C_\mu \inf_{m\in\mathcal{M}} \left\{\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + pen(m)\right\} + \frac{C(\rho, \mu, \Gamma, \Sigma)}{n}. \tag{4.3}$$

This theorem provides a non asymptotic bound for the Hellinger risk between the selected model and the true one. On the opposite of Theorem 3.1, the penalty function does not depend on the bound of the true function. The selection procedure based only on the data offers the advantage to free the estimator from any prior knowledge about the smoothness of the function to estimate. The estimator is therefore adaptive. As we bound Hellinger risk in (4.3) by

Kulback-Leibler risk, one should prefer to have the Hellinger risk on the right hand side instead of the Kulback-Leibler risk. Such a bound is possible if we assume that $\log(\|\pi_{f_0}/\rho\|_\infty)$ is bounded. Indeed if we assume that there exists T such that $\log(\|\pi_{f_0}/\rho\|_\infty) \leq T$, this implies that $\log(\|\pi_{f_0}/\pi_{f_m}\|_\infty) \leq T$ uniformly for all partitions $m \in \mathcal{M}$. Now using Inequality (7.6) p. 362 in Birgé and Massart (1998) we have that $\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) \leq (4 + 2\log(M))h^2(\mathbb{P}_{f_0}, \mathbb{P}_{f_m})$ which implies,

$$\mathbb{E}_{f_0}[h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\tilde{f}})] \leqslant C_\mu.C(T) \inf_{m\in\mathcal{M}} \Big\{h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\Big\} + \frac{C(\rho,\mu,\Gamma,\Sigma)}{n}.$$

Choice of the weights $\{L_m, m \in \mathcal{M}\}$. According to Theorem 4.1, the penalty function depends on the collection $\mathcal{M}$ through the choice of the weights L_m satisfying (4.2), *i.e.*

$$\Sigma = \sum_{m\in -M} \exp(-L_m D_m) = \sum_{D\geq 1} e^{-L_D D} Card\{m \in \mathcal{M}, |m| = D\} < \infty. \tag{4.4}$$

Hence the number of models having the same dimension D plays an important role in the risk bound.

If there is only one model of dimension D, a simple way of choosing L_D is to take them constant, *i.e.* $L_D = L$ for all $m \in \mathcal{M}$, and thus we have from (4.4)

$$\Sigma = \sum_{D\geq 1} e^{-LD} < \infty.$$

This is the case when $\mathcal{M}$ is a family of regular partitions. Consequently, the choice *i.e.* $L_D = L$ for all $m \in \mathcal{M}$ leads to a penalty proportional to the dimension D_m, and for every $D_m \geq 1$,

$$\mathrm{pen}(m) = \mu\Big(1 + 6L + 8\sqrt{L}\Big)\frac{D_m}{n} = c \times \frac{D_m}{n}. \tag{4.5}$$

In the more general context, that is in the case of irregular partitions, the numbers of models having the same dimension D is exponential and satisfies

$$Card\Big\{m \in \mathcal{M}, |m| = D\Big\} = \binom{n-1}{D-1} \leq \binom{n}{D}.$$

In that case we choose L_m depending on the dimension D_m. With L depending on D, Σ in (4.2) satisfies

$$\begin{aligned}
\Sigma &= \sum_{D\geq 1} e^{-L_D D} Card\{m \in \mathcal{M}, |m| = D\} \\
&\leq \sum_{D\geq 1} e^{-L_D D} \binom{n}{D} \\
&\leq \sum_{D\geq 1} e^{-L_D D} \Big(\frac{en}{D}\Big)^D \\
&\leq \sum_{D\geq 1} e^{-D\Big(L_D - 1 - \log\left(\frac{n}{D}\right)\Big)}
\end{aligned}$$

So taking $L_D = 2 + \log(\frac{n}{D})$ leads to $\Sigma < \infty$ and the penalty becomes

$$\mathrm{pen}(m) = \mu \times \mathrm{pen}_{\text{shape}}(m), \tag{4.6}$$

where

$$\mathrm{pen}_{\text{shape}}(m) = \frac{D_m}{n}\Big[13 + 6\log\Big(\frac{n}{D_m}\Big) + 8\sqrt{2 + \log\Big(\frac{n}{D_m}\Big)}\Big]. \tag{4.7}$$

The constant μ can be calibrated using the slope heuristics Birgé and Massart (2007) (see Section 5.2).

Remark 4.1. *In Theorem 4.1, we do not assume that the target function f_0 is piecewise constant. However in many contexts, for instance in segmentation, we might want to consider that f_0 is piecewise constant or can be well approximated by piecewise constant functions. That means there exists of partition of $\mathcal{X}$ within which the observations follow the same distribution and between which observations have different distributions.*

5. Simulations

In this section we present numerical simulation to study the non-asymptotic properties of the model selection procedure introduced in Section 4.4. More precisely, the numerical properties of the estimators built by model selection with our criteria are compared with those of the estimators resulting from model selection using the well known criteria AIC and BIC.

5.1. **Simulations frameworks.** We consider the model defined in (2.1) with $f_0 : [0,1] \to \mathbb{R}$. The aim is to estimate f_0. We consider the collection of models $(S_m)_{m\in\mathcal{M}}$, where

$$S_m = \mathrm{Vect}\{\mathbb{1}_{[\frac{k-1}{D_m},\frac{k}{D_m}[} \text{ such that } 1 \le k \le D_m\},$$

and $\mathcal{M}$ is the collection of regular partitions

$$m = \left\{\Big[\frac{k-1}{D_m}, \frac{k}{D_m}\Big[, \text{ such that } 1 \le k \le D_m, \right\},$$

where

$$D_m \le \frac{n}{\log n}.$$

The collection of estimators is defined in Lemma 4.1. Let us thus consider four penalties.

- the AIC criretion defined by

$$\mathrm{pen}_{\text{AIC}} = \frac{D_m}{n};$$

- the BIC criterion defined by

$$\mathrm{pen}_{\text{BIC}} = \frac{\log n}{2n} D_m;$$

- the penalty proportional to the dimension as in (4.5) defined by

$$\text{pen}_{\text{lin}} = c \times \frac{D_m}{n};$$

- and the penalty defined in (4.6) by

$$\text{pen} = \mu \times \text{pen}_{\text{shape}}(m).$$

pen_{lin} and pen are penalties depending on some unknown multiplicative constant (c and μ respectively) to be calibrated. As previously said we will use the "slope heuristics" introduced in Birgéa nd Massart (2007) to calibrate the multiplicative constant. We have distinguished two cases:

- The case where there exists $m_o \in \mathcal{M}$ such that the true function belong to S_{m_o} *i.e.* where f_0 is piecewise constant,

$$\begin{aligned} \text{Mod1: } f_0 &= 0.5\,\mathbb{1}_{[0,1/3)} + \mathbb{1}_{[1/3,0.5)} + 2\,\mathbb{1}_{[0.5,2/3)} + 0.25\,\mathbb{1}_{[2/3,1]} \\ \text{Mod2: } f_0 &= 0.75\,\mathbb{1}_{[0,1/4]} + 0.5\,\mathbb{1}_{[1/4,0.5)} + 0.2\,\mathbb{1}_{[0.5,3/4)} + 0.3\,\mathbb{1}_{[3/4,1]}. \end{aligned}$$

- The second case, f_0 does not belong to any S_m, $m \in \mathcal{M}$ and is chosen in the following way:

$$\begin{aligned} \text{Mod3: } f_0(x) &= \sin(\pi x) \\ \text{Mod4: } f_0(x) &= \sqrt{x}. \end{aligned}$$

In each case, the x_i's are simulated according to uniform distribution on $[0, 1]$.

The Kullback-Leibler divergence is definitely not suitable to evaluate the quality of an estimator. Indeed, given a model S_m, there is a positive probability that on one of the interval $I \in m$ we have $\pi^{(I)}_{\hat{f}_m} = 0$ or $\pi^{(I)}_{\hat{f}_m} = 1$, which implies that $\mathcal{K}(\pi^{(n)}_{f_0}, \pi^{(n)}_{\hat{f}_m}) = +\infty$. So we will use the Hellinger distance to evaluate the quality of an estimator.

Even if an oracle inequality seems of no practical use, it can serve as a benchmark to evaluate the performance of any data driven selection procedure. Thus model selection performance of each procedure is evaluated by the following benchmark

$$C^* := \frac{\mathbb{E}\Big[h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\Big]}{\mathbb{E}\Big[\inf_{m\in\mathcal{M}} h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m})\Big]}. \tag{5.8}$$

C^* evaluate how far is the selected estimator to the oracle. The values of C^* evaluated for each procedure with different sample size $n \in \{100, 200, \ldots, 1000\}$ are reported in Figure 2 , Figure 4, Figure 3 and Figure 5. For each sample size $n \in \{100, 200, \ldots, 1000\}$, the expectation was estimated using mean over 1000 simulated datasets.

5.2. **Slope heuristics.** The aim of this section is to show how the penalty in Theorem 4.1 can be calibrated in practice using the main ideas of data-driven penalized model selection criterion proposed by Birgé and Massart (2007). We calibrate penalty using "slope heuristics" first introduced and theoretically validated by Birgé and Massart (2007) in a gaussian homoscedastic setting. Recently it has also been theoretically validated in the heteroscedastic random-design case by Arlot (2009) and for least squares density estimation by Lerasle (2012). Several encouraging applications of this method are developed in many other frameworks (see for instance in clustering and variable selection for categorical multivariate data Bontemps and Toussile (2013), for variable selection and clustering via Gaussian mixtures Maugis and Michel (2011), in multiple change points detection Lebarbier (2005)). Some overview and implementation of the slope heuristics can be find in Baudry *et al.* (2012).

We now describe the main idea of those heuristics, starting from that main goal of the model selection, that is to choose the best estimator of f_0 among a collection of estimators $\{\hat{f}_m\}_{m\in\mathcal{M}}$. Moreover, we expect that this best estimator mimics the so-called oracle defined as (2.7). To this aim, the great challenge is to build a penalty function such that the selected model $\hat{m}$ is nearly as good as the oracle. In the following we call the ideal penalty the penalty that leads to the choice of $m*$. Using that

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m}) = \gamma(\hat{f}_m) - \gamma(f_0),$$

then, by definition, $m*$ defined in (2.7) satisfies

$$m* = \arg\min_{m\in\mathcal{M}}[\gamma(\hat{f}_m) - \gamma(f_0)] = \arg\min_{m\in\mathcal{M}} \gamma(\hat{f}_m).$$

The ideal penalty, leading to the choice of the oracle $m*$, is thus $[\gamma(\hat{f}_m) - \gamma_n(\hat{f}_m)]$, for $m \in \mathcal{M}$. As the matter of fact, by replacing $\mathrm{pen}_{id}(\hat{f}_m)$ by its value, we obtain

$$\begin{aligned}\arg\min_{m\in\mathcal{M}}[\gamma_n(\hat{f}_m) + \mathrm{pen}_{id}(\hat{f}_m)] &= \arg\min_{m\in\mathcal{M}}[\gamma_n(\hat{f}_m) + \gamma(\hat{f}_m) - \gamma_n(\hat{f}_m)] \\ &= \arg\min_{m\in\mathcal{M}}[\gamma(\hat{f}_m)] \\ &= m*.\end{aligned}$$

Of course this ideal penalty always selects the oracle model but depends on the unknown function f_0 throught the sample distribution, since $\gamma(t) = \mathbb{E}_{f_0}[\gamma_n(t)]$. A natural idea is to choose $\mathrm{pen}(m)$ as close as possible to $\mathrm{pen}_{id}(m)$ for every $m \in \mathcal{M}$. Now, we use that this ideal penalty can be decomposed into

$$\mathrm{pen}_{id}(m) = \gamma(\hat{f}_m) - \gamma_n(\hat{f}_m) = v_m + \hat{v}_m + e_m,$$

where

$$v_m = \gamma(\hat{f}_m) - \gamma(f_m), \quad \hat{v}_m = \gamma_n(f_m) - \gamma_n(\hat{f}_m), \text{ and } \quad e_m = \gamma(f_m) - \gamma_n(f_m).$$

The slope heuristics relies on two points:

- The existence of a minimal penalty $\mathrm{pen}_{\min}(m) = \hat{v}_m$ such that when the penalty is smaller than $\mathrm{pen}_{\min}$ the selected model is one of the most complex models. Whereas,

penalties larger than pen_{min} lead to a selection of models with "reasonable" complexity.

- Using concentration arguments, it is reasonable to consider that uniformly over $\mathcal{M}$, $\gamma_n(f_m)$ is close to its expectation which implies that $e_m \approx 0$. In the same way, since $\hat{v}_m$ is a empirical version of v_m, it is also reasonable to consider that $v_m \approx \hat{v}_m$. Ideal penalty is thus approximately given by $2\hat{v}_m$, and thus

$$\text{pen}_{id}(m) \quad \approx \quad 2\text{pen}_{min}(m).$$

In practice, $\hat{v}_m$ can be estimated from the data provided that ideal penalty $\text{pen}_{id}(.) = \kappa_{id}\text{pen}_{shape}(.)$ is known up to a multiplicative factor. A major point of the slope heuristics is that

$$\frac{\kappa_{id}}{2}\text{pen}_{shape}(.)$$

is a good estimator of $\hat{v}_m$ and this provides the minimal penalty.

Provided that $\text{pen} = \kappa \times \text{pen}_{shape}$ is known up to a multiplicative constant κ that is to be calibrated, we combine the previously heuristic to the method usually known as dimension jump method. In practice, we consider a grid $\kappa_1, \ldots, \kappa_M$, where each κ_j leads to a selected model $\hat{m}_{\kappa_i}$ with dimension $D_{\hat{m}_{\kappa_i}}$. The constant κ_{min} which corresponds to the value such that $\text{pen}_{min} = \kappa_{min} \times \text{pen}_{shape}$, is estimated using the first point of the "slope heuristics". If $D_{\hat{m}_{\kappa_j}}$ is plotted as a function of κ_j, κ_{min} is such that $D_{\hat{m}_{\kappa_j}}$ is "huge" for $\kappa < \kappa_{min}$ and "reasonably small" for $\kappa > \kappa_{min}$. So κ_{min} is the value at the position of the biggest jump. For more details about this method we refer the reader to Baudry *et al.* (2012) and Arlot and Massart (2009).

Figures 2 and 3 are the cases where the true function is piecewise constant. Figure 4 and Figure 5 are situations where the true function does not belong to any model in the given collection. The performance of criteria depends on the sample size n. In these two situations we observe that our two model selection procedures are comparable, and their performance increases with n. While the performance of model selected by BIC decreases with n. Our criteria outperformed the AIC for all n. The BIC criterion is better than our criteria for $n \leq 200$. For $200 < n \leq 400$, the performance of the model selected by BIC is quite the same as the performance of models selected by our criteria. Finally for $n > 400$ our criteria outperformed the BIC.

Theoretical results and simulations raise the following question : why our criteria are better than BIC for quite large values of n yet theoretical results are non asymptotic? To answer this question we can say that, in simulations, to calibrate our penalties we have used "slope heuristics", and those heuristic are based on asymptotic arguments (see Section 5.2).

6. Proofs

6.1. **Notations and technical tools.** Subsequently we will use the following notations. Denote by $\| f \|_n$ and $\langle f, g\rangle_n$ the empirical euclidian norm and the inner product

$$\| f \|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f^2(x_i), \text{ and } \langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i).$$

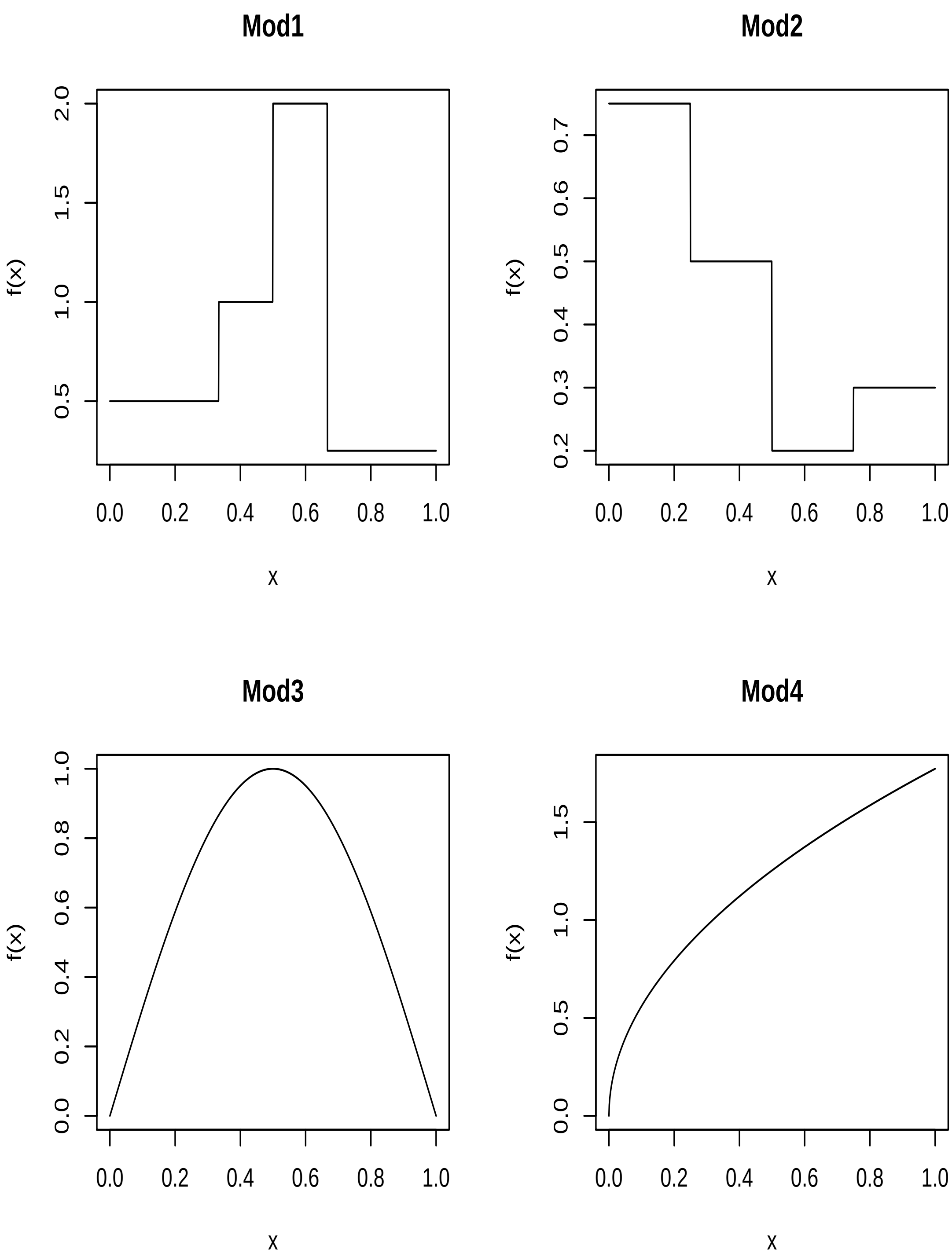

FIGURE 1. Different functions f_0 to be estimated

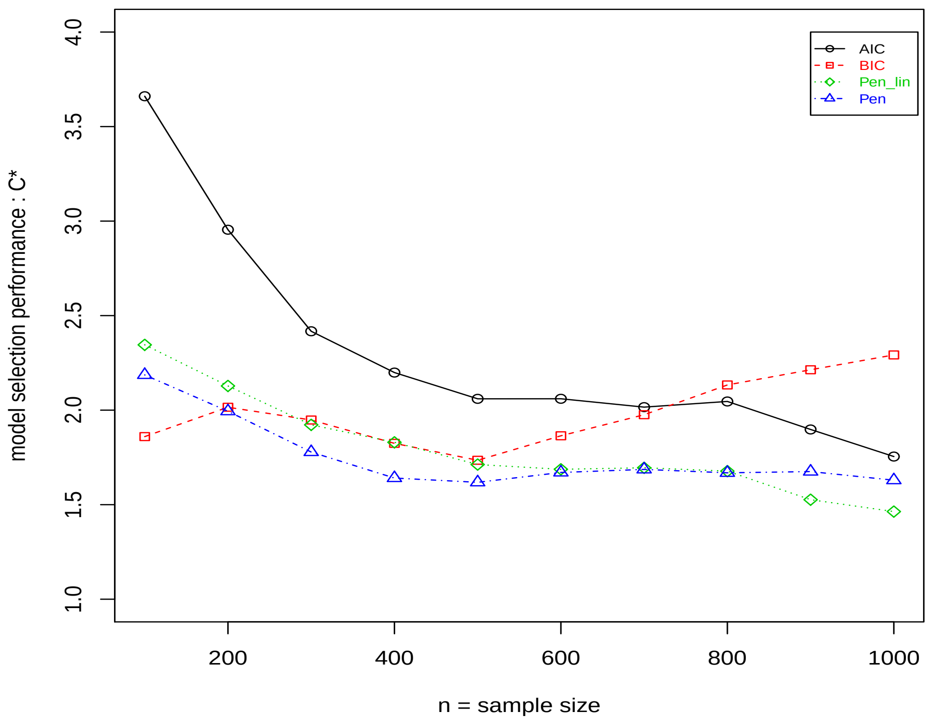

FIGURE 2. Model selection performance (C^*) as a function of sample size n, with each penalty, Mod1.

Note that $\| \,.\, \|_n$ is a semi norm on the space $\mathcal{F}$ of functions $g : \mathcal{X} \longrightarrow \mathbb{R}$, but is a norm in the quotient space $\mathcal{F}/\mathcal{R}$ associated to the equivalence relation $\mathcal{R}$: $g \,\mathcal{R}\, h$ if and only if $g(x_i) = h(x_i)$ for all $i \in \{1, \ldots, n\}$. It follows from (2.3) that γ defined in (2.4) can be expressed as the sum of a centered empirical process and of the estimation criterion γ_n. More precisely, denoting by $\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$, with $\varepsilon_i = Y_i - \mathbb{E}_{f_0}(Y_i)$, for all f, we have

$$\gamma(f) = \gamma_n(f) + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) := \gamma_n(f) + \langle \vec{\varepsilon}, f \rangle_n. \tag{6.1}$$

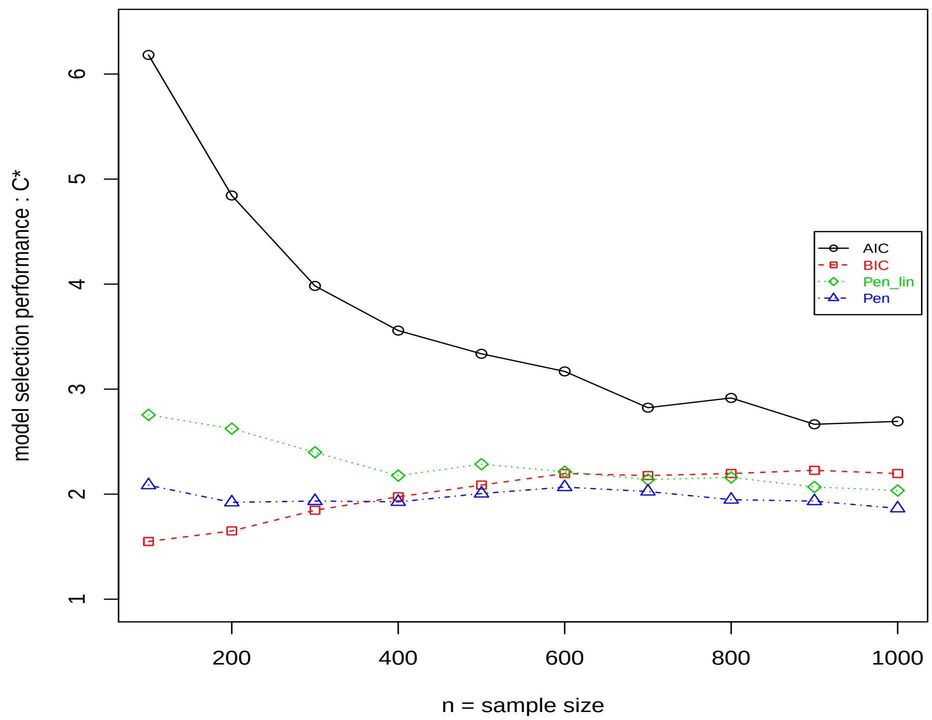

FIGURE 3. Model selection performance (C^*) as a function of sample size n, with each penalty, Mod2.

Easy calculations show that for γ defined in (2.4) we have,

$$\begin{aligned}\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f}) &= \frac{1}{n}\int \log\left(\frac{\mathbb{P}^{(n)}_{f_0}}{\mathbb{P}^{(n)}_{f}}\right) d\mathbb{P}^{(n)}_{f_0} = \gamma(f) - \gamma(f_0) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\pi_{f_0}(x_i)\log\left(\frac{\pi_{f_0}(x_i)}{\pi_f(x_i)}\right) + (1-\pi_{f_0}(x_i))\log\left(\frac{1-\pi_{f_0}(x_i)}{1-\pi_f(x_i)}\right)\right].\end{aligned}$$

Let us recall the usual bounds (see Castellan (2003b)) for kullback-Leibler information:

Lemma 6.1. *For positive densities p and q with respect to μ, if $f = \log(q/p)$, then*

$$\frac{1}{2}\int f^2(1\wedge e^f)p\,d\mu \leqslant \mathcal{K}(p,q) \leqslant \frac{1}{2}\int f^2(1\vee e^f)p\,d\mu.$$

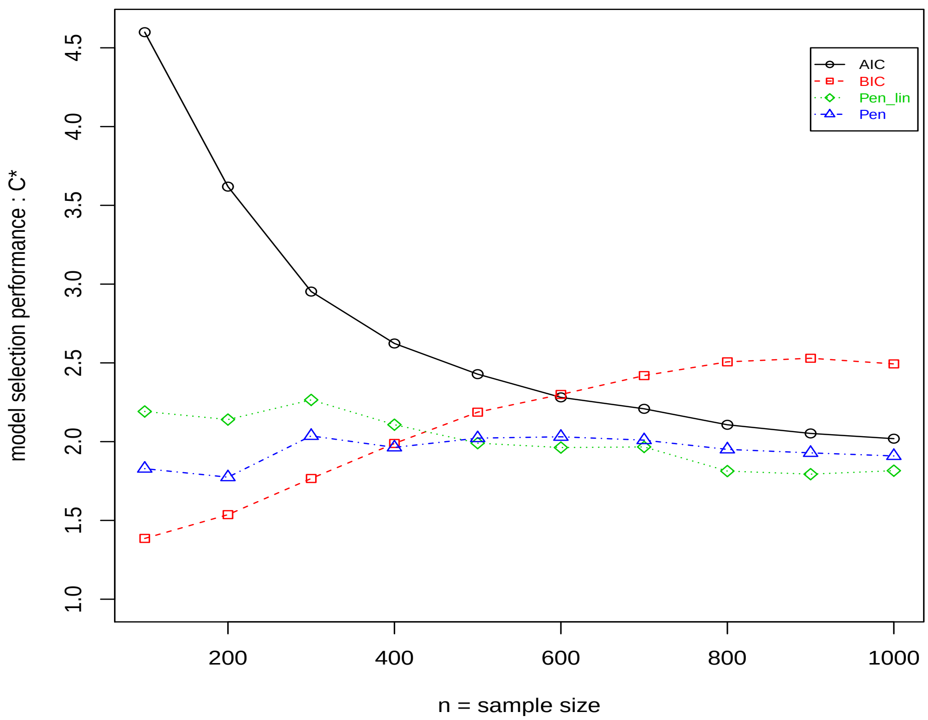

FIGURE 4. Model selection performance (C^*) as a function of sample size n, with each penalty, Mod3.

6.2. **Proof of Proposition 3.1:** By definition of $\hat{f}_m$, for all $f \in S_m \cap \mathbb{L}_\infty(C_0)$, $\gamma_n(\hat{f}_m) - \gamma_n(f) \leqslant 0$. We apply (6.1), with $f = f_m$ and $f = \hat{f}_m$,

$$\gamma(\hat{f}_m) - \gamma(f_0) \leqslant \gamma(f_m) - \gamma(f_0) + \langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n.$$

As usual, the main part of the proof relies on the study of the empirical process $\langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n$. Since $\hat{f}_m - f_m$ belongs to S_m, $\hat{f}_m - f_m = \sum_{j=1}^{D_m} \alpha_j \psi_j$, where $\{\psi_1, \ldots, \psi_{D_m}\}$, is an orthonormal basis of S_m and consequently

$$\langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n = \sum_{j=1}^{D_m} \alpha_j \langle \vec{\varepsilon}, \psi_j \rangle_n.$$

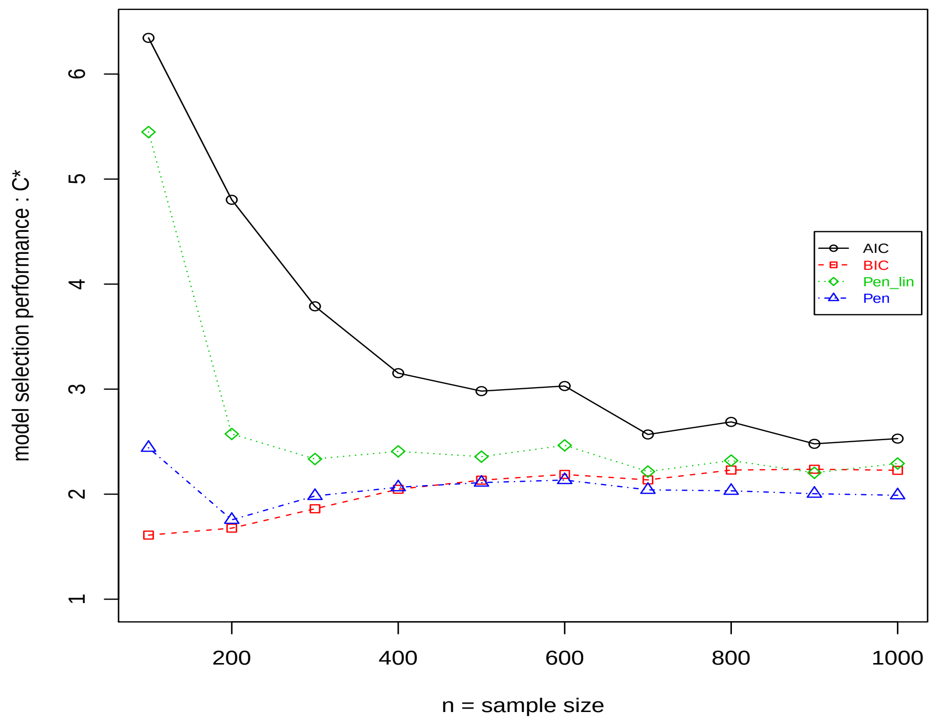

FIGURE 5. Model selection performance (C^*) as a function of sample size n, with each penalty, Mod4.

Applying Cauchy-Schwarz inequality we get

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n &\leqslant \sqrt{\sum_{j=1}^{D_m} \alpha_j^2} \sqrt{\sum_{j=1}^{D_m} \left(\langle \vec{\varepsilon}, \psi_j \rangle_n \right)^2} \\
&= \|\hat{f}_m - f_m\|_n \sqrt{\sum_{j=1}^{D_m} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_j(x_i) \right)^2}.
\end{aligned}$$

We now apply Lemma 6.2 (See Section 7 for the proof of Lemma 6.2)

Lemma 6.2. *Let $\mathcal{S}_m$ the model defined in (2.5) and $\{\psi_1, \dots, \psi_{D_m}\}$ an orthonormal basis of the linear span $\{\phi_k,\ k \in m\}$. We also denote by Λ_m the set of $\beta = (\beta_1, ..., \beta_D)$ such that $f_\beta(.) = \sum_{j=1}^{D} \beta_j \psi_j(.)$ satisfies $f_\beta \in \mathcal{S}_m \cap \mathbb{L}_\infty(C_0)$. Let β^* be any minimizer of the function $\beta \to \gamma(f_\beta)$ over*

Λ_m, *we have*

$$\frac{\mathcal{U}_0^2}{2}\|f_\beta - f_{\beta^*}\|_n^2 \leq \gamma(f_\beta) - \gamma(f_{\beta^*}), \tag{6.2}$$

where $\mathcal{U}_0 = e^{C_0}/(1+e^{C_0})^2$.

Then we have

$$\langle \vec{\varepsilon}, \hat{f}_m - f_m\rangle_n \quad \leqslant \quad \sqrt{\sum_{j=1}^{D_m}\Big(\langle \vec{\varepsilon}, \psi_j\rangle_n\Big)^2}\,\frac{\sqrt{2}}{\mathcal{U}_0}\sqrt{\gamma(\hat{f}_m) - \gamma(f_m)}$$

Now we use that for every positive numbers, a, b, x, $ab \leqslant (x/2)a^2 + [1/(2x)]b^2$, and infer that

$$\gamma(\hat{f}_m) - \gamma(f_0) \leq \gamma(f_m) - \gamma(f_0) + \frac{x}{\mathcal{U}_0^2}\sum_{j=1}^{D_m}\Big(\langle \vec{\varepsilon}, \psi_j\rangle_n\Big)^2 + (1/2x)(\gamma(\hat{f}_m) - \gamma(f_m)).$$

For $x > 1/2$, it follows that

$$\mathbb{E}_{f_0}[\gamma(\hat{f}_m) - \gamma(f_0)] \leqslant \gamma(f_m) - \gamma(f_0) + \frac{2x^2}{(2x-1)\mathcal{U}_0^2}\mathbb{E}_{f_0}\left[\sum_{j=1}^{D_m}\Big(\langle \vec{\varepsilon}, \psi_j\rangle_n\Big)^2\right].$$

We conclude the proof by using that

$$\mathbb{E}_{f_0}\left[\sum_{j=1}^{D_m}\Big(\langle \vec{\varepsilon}, \psi_j\rangle_n\Big)^2\right] \leqslant \frac{D_m}{4n}.$$

□

6.3. **Proof of Theorem 3.1.** By definition, for all $m \in \mathcal{M}$,

$$\gamma_n(\hat{f}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leqslant \gamma_n(\hat{f}_m) + \mathrm{pen}(m) \leqslant \gamma_n(f_m) + \mathrm{pen}(m).$$

Applying (6.1) we have

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m\rangle_n + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}). \tag{6.3}$$

It remains to study $\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m\rangle_n$, using the following lemma, which is a modification of Lemma 1 in Durot *et al.* (2009).

Lemma 6.3. *For every D, D' and $x \geqslant 0$ we have*

$$\mathbb{P}\left(\sup_{u\in\left(S_D\cap\mathbb{L}_\infty(C_0)+S_{D'}\cap\mathbb{L}_\infty(C_0)\right)}\frac{\langle \vec{\varepsilon}, u\rangle_n}{\| u \|_n}\rangle \sqrt{\frac{D+D'}{4n}} + \sqrt{\frac{5x}{n}}\right) \leqslant \exp(-x).$$

Fix $\xi > 0$ and let $\Omega_\xi(m)$ denote the event

$$\Omega_\xi(m) = \bigcap_{m' \in \mathcal{M}} \left\{ \sup_{u \in \left(S_m \cap \mathbb{L}_\infty(C_0) + S_{m'} \cap \mathbb{L}_\infty(C_0)\right)} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} \leq \sqrt{\frac{D_m + D_{m'}}{4n}} + \sqrt{5(L_{m'} D_{m'} + \xi)/n} \right\}.$$

Then we have

$$\mathbb{P}\left(\Omega_\xi(m)\right) \geqslant 1 - \Sigma \exp(-\xi). \tag{6.4}$$

See the Appendix for the proof of this lemma. Fix $\xi > 0$, applying Lemma 6.3, we infer that on the event $\Omega_\xi(m)$,

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n &\leqslant \left(\sqrt{\frac{D_m + D_{\hat{m}}}{4n}} + \sqrt{5 \frac{L_{\hat{m}} D_{\hat{m}} + \xi}{n}} \right) \| \hat{f}_{\hat{m}} - f_m \|_n \\
&\leqslant \left(\sqrt{\frac{D_m + D_{\hat{m}}}{4n}} + \sqrt{5 \frac{L_{\hat{m}} D_{\hat{m}} + \xi}{n}} \right) \left(\| \hat{f}_{\hat{m}} - f_0 \|_n + \| f_0 - f_m \|_n \right) \\
&\leqslant \left(\sqrt{D_{\hat{m}}} \left(\frac{1}{\sqrt{4n}} + \sqrt{\frac{5 L_{\hat{m}}}{n}} \right) + \sqrt{\frac{D_m}{4n}} + \sqrt{5 \frac{\xi}{n}} \right) \left(\| \hat{f}_{\hat{m}} - f_0 \|_n + \| f_0 - f_m \|_n \right).
\end{aligned}$$

Applying that $2xy \leqslant \theta x^2 + \theta^{-1} y^2$, for all $x > 0$, $y > 0$, $\theta > 0$, we get that on $\Omega_\xi(m)$ and for every $\eta \in]0, 1[$

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n &\leqslant (\frac{1 - \eta}{2}) \left[(1 + \eta) \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + (1 + \eta^{-1}) \| f_0 - f_m \|_n^2 \right] \\
&+ \frac{1}{2(1 - \eta)} \left[(1 + \eta) D_{\hat{m}} \left(\frac{1}{\sqrt{4n}} + \sqrt{\frac{5 L_{\hat{m}}}{n}} \right)^2 + (1 + \eta^{-1}) \left(\sqrt{\frac{D_m}{4n}} + \sqrt{\frac{5\xi}{n}} \right)^2 \right] \\
&\leqslant \frac{1 - \eta^2}{2} \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + \frac{\eta^{-1} - \eta}{2} \| f_0 - f_m \|_n^2 + \frac{1 + \eta}{2(1 - \eta)} D_{\hat{m}} \left(\frac{1}{\sqrt{4n}} + \sqrt{\frac{5 L_{\hat{m}}}{n}} \right)^2 \\
&\quad + \frac{1 + \eta^{-1}}{1 - \eta} (\frac{D_m}{4n} + \frac{5\xi}{n}).
\end{aligned}$$

If $\text{pen}(m) \geqslant \left(\lambda D_m \left(\frac{1}{2} + \sqrt{5 L_m} \right)^2 \right) / n$, with $\lambda > 0$, we have

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n \leqslant{}& \frac{1 - \eta^2}{2} \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + \frac{\eta^{-1} - \eta}{2} \| f_0 - f_m \|_n^2 + \frac{1 + \eta}{2(1 - \eta)\lambda} \text{pen}(\hat{m}) + \frac{1 + \eta^{-1}}{(1 - \eta)\lambda} \text{pen(m)} \\
&+ \frac{1 + \eta^{-1}}{1 - \eta} \frac{5\xi}{n}.
\end{aligned}$$

It follows from (6.3) that

$$\begin{aligned}\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leqslant\ & \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) + \frac{1-\eta^2}{2} \parallel \hat{f}_{\hat{m}} - f_0 \parallel_n^2 + \frac{\eta^{-1}-\eta}{2} \parallel f_0 - f_m \parallel_n^2 \\ & + \frac{1+\eta}{2(1-\eta)\lambda}\mathrm{pen}(\hat{m}) + \frac{1+\eta^{-1}}{(1-\eta)\lambda}\mathrm{pen(m)} + \frac{1+\eta^{-1}}{1-\eta}\frac{5\xi}{n} + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}).\end{aligned}$$

Taking $\lambda = (\eta+1)/(2(1-\eta))$, we have

$$\begin{aligned}&\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) \\ &\quad + \frac{4\lambda}{(2\lambda+1)^2} \parallel \hat{f}_{\hat{m}} - f_0 \parallel_n^2 + \frac{4\lambda}{4\lambda^2-1} \parallel f_0 - f_m \parallel_n^2 + \frac{6\lambda+1}{2\lambda-1}\mathrm{pen}(m) + \frac{10\lambda(2\lambda+1)}{2\lambda-1}\frac{\xi}{n}.\end{aligned}$$

Now we use the following lemma (see Lemma 6.1 in Kwemou (2012)) that allows to connect empirical norm and Kullback-Leibler divergence.

Lemma 6.4. *Under Assumptions* ($\mathbf{A_1}$)*, for all* $m \in \mathcal{M}$ *and all* $t \in S_m \cap \mathbb{L}_\infty(C_0)$*, we have*

$$c_{min}\|t-f_0\|_n^2 \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{t}) \leqslant c_{max}\|t-f_0\|_n^2.$$

where c_{min} *and* c_{max} *are constants depending on* C_0 *and* c_1.

Consequently

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leqslant C(c_{min})\left\{\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\right\} + C_1(c_{min})\frac{\xi}{n},$$

where

$$C(c_{min}) = \max\left\{\frac{1+\frac{4\lambda}{(4\lambda^2-1)c_{min}}}{1-\frac{4\lambda}{c_{min}(2\lambda+1)^2}}; \frac{\frac{6\lambda+1}{2\lambda-1}}{1-\frac{4\lambda}{c_{min}(2\lambda+1)^2}}\right\} \text{ and } C_1(c_{min}) = \frac{\frac{10\lambda(2\lambda+1)}{2\lambda-1}}{1-\frac{4\lambda}{c_{min}(2\lambda+1)^2}}.$$

Thus we take λ such that

$$1-\frac{4\lambda}{c_{min}(2\lambda+1)^2} > 0, \tag{6.5}$$

where c_{min} depends on the bound of the true function f_0. By definition of $\Omega_\xi(m)$ and (6.4), there exists a random variable $V \geqslant 0$ with $\mathbb{P}(V>\xi) \leqslant \Sigma\exp(-\xi)$ and $\mathbb{E}_{f_0}(V) \leqslant \Sigma$, such that

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leqslant C(c_{min})\left\{\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\right\} + C_1(c_{min})\frac{V}{n},$$

which implies that for all $m \in \mathcal{M}$,

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})] \leqslant C(c_{min})\left\{\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\right\} + C_1(c_{min})\frac{\Sigma}{n}.$$

This concludes the proof. □

6.4. **Proof of Proposition 4.1:** Let f_m, $\hat{f}_m$, π_{f_m} and $\pi_{\hat{f}_m}$ given in Lemma 4.1, proved in appendix. In the following, $D_m = |m|$. For $\delta > 0$, let $\Omega_m(\delta)$ be the event

$$\Omega_m(\delta) = \bigcap_{J\in m} \left\{ \left| \frac{\pi^{(J)}_{\hat{f}_m}}{\pi^{(J)}_{f_m}} - 1 \right| \leqslant \delta \right\} \bigcap \left\{ \left| \frac{1-\pi^{(J)}_{\hat{f}_m}}{1-\pi^{(J)}_{f_m}} - 1 \right| \leqslant \delta \right\}. \tag{6.6}$$

According to pythagore's type identity and Lemma 4.1 we write

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m}) = \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathcal{K}(\mathbb{P}^{(n)}_{f_m}, \mathbb{P}^{(n)}_{\hat{f}_m}) \mathbb{1}_{\Omega_m(\delta)} + \mathcal{K}(\mathbb{P}^{(n)}_{f_m}, \mathbb{P}^{(n)}_{\hat{f}_m}) \mathbb{1}_{\Omega^c_m(\delta)},$$

where

$$\begin{aligned} \mathcal{K}(\mathbb{P}^{(n)}_{f_m}, \mathbb{P}^{(n)}_{\hat{f}_m}) &= \frac{1}{n}\sum_{i=1}^{n} \left[\pi_{f_m}(x_i) \log\left(\frac{\pi_{f_m}(x_i)}{\pi_{\hat{f}_m}(x_i)} \right) + (1-\pi_{f_m}(x_i)) \log\left(\frac{1-\pi_{f_m}(x_i)}{1-\pi_{\hat{f}_m}(x_i)} \right) \right] \\ &= \frac{1}{n}\sum_{J\in m} |J| \left[\pi^{(J)}_{f_m} \log\left(\frac{\pi^{(J)}_{f_m}}{\pi^{(J)}_{\hat{f}_m}} \right) + (1-\pi^{(J)}_{f_m}) \log\left(\frac{1-\pi^{(J)}_{f_m}}{1-\pi^{(J)}_{\hat{f}_m}} \right) \right]. \end{aligned} \tag{6.7}$$

The first step consists in showing that

$$\frac{1-\delta}{2(1+\delta)^2} \mathcal{X}^2_m \mathbb{1}_{\Omega_m(\delta)} \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_m}, \mathbb{P}^{(n)}_{\hat{f}_m}) \mathbb{1}_{\Omega_m(\delta)} \leqslant \frac{1+\delta}{2(1-\delta)^2} \mathcal{X}^2_m \mathbb{1}_{\Omega_m(\delta)}, \tag{6.8}$$

where

$$\mathcal{X}^2_m = \frac{1}{n}\sum_{J\in m} \frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]}, \text{ with } \qquad \frac{4\rho^2 D_m}{n} \leqslant \mathbb{E}_{f_0}[\mathcal{X}^2_m] \leqslant \frac{2D_m}{n}. \tag{6.9}$$

The second step relies on the proof of

$$\left| \mathbb{E}_{f_0}\left(\mathcal{K}(\mathbb{P}^{(n)}_{f_m}, \mathbb{P}^{(n)}_{\hat{f}_m}) \mathbb{1}_{\Omega^c_m(\delta)} \right) \right| \leqslant 2\log\left(\frac{1}{\rho}\right) \mathbb{P}[\Omega^c_m(\delta)]. \tag{6.10}$$

The last step consists in showing that for $\epsilon > 0$, since for all $J \in m$, $|J| \geq \Gamma[\log(n)]^2$, where $\Gamma > 0$ is an absolute constant, then we have

$$\mathbb{P}[\Omega^c_m(\delta)] \leqslant 4|m| \exp\left(-\frac{\delta^2}{2(1+\delta/3)} \rho^2 \Gamma[\log(n)]^2 \right) \leq \frac{\kappa(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}}. \tag{6.11}$$

Gathering (6.8)-(6.11), we conclude that

$$\begin{aligned} \mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m})] &\leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \frac{(1+\delta)|m|}{(1-\delta)^2 n} + 2\log\left(\frac{1}{\rho}\right) \mathbb{P}[\Omega^c_m(\delta)] \\ &\leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \frac{(1+\delta)|m|}{(1-\delta)^2 n} + \frac{\kappa(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}}. \end{aligned}$$

We finish by proving (6.8), (6.9), (6.10) and (6.11).

• Proof of (6.8) and (6.9) : Arguing as in Castellan (2003b) and using Lemma 6.1 we have

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_m},\mathbb{P}^{(n)}_{\hat f_m}) \geqslant \frac{1}{2n}\sum_{J\in m}|J|\left[\pi^{(J)}_{f_m}\left(1\wedge\frac{\pi^{(J)}_{\hat f_m}}{\pi^{(J)}_{f_m}}\right)\log^2\left(\frac{\pi^{(J)}_{f_m}}{\pi^{(J)}_{\hat f_m}}\right)+(1-\pi^{(J)}_{f_m})\left(1\wedge\frac{1-\pi^{(J)}_{\hat f_m}}{1-\pi^{(J)}_{f_m}}\right)\log^2\left(\frac{1-\pi^{(J)}_{f_m}}{1-\pi^{(J)}_{\hat f_m}}\right)\right]$$

and

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_m},\mathbb{P}^{(n)}_{\hat f_m}) \leqslant \frac{1}{2n}\sum_{J\in m}|J|\left[\pi^{(J)}_{f_m}\left(1\vee\frac{\pi^{(J)}_{\hat f_m}}{\pi^{(J)}_{f_m}}\right)\log^2\left(\frac{\pi^{(J)}_{f_m}}{\pi^{(J)}_{\hat f_m}}\right)+(1-\pi^{(J)}_{f_m})\left(1\vee\frac{1-\pi^{(J)}_{\hat f_m}}{1-\pi^{(J)}_{f_m}}\right)\log^2\left(\frac{1-\pi^{(J)}_{f_m}}{1-\pi^{(J)}_{\hat f_m}}\right)\right].$$

It follows that

$$\frac{1-\delta}{2}V^2(\pi_{f_m},\pi_{\hat f_m})\mathbb{1}_{\Omega_m(\delta)} \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_m},\mathbb{P}^{(n)}_{\hat f_m})\mathbb{1}_{\Omega_m(\delta)} \leqslant \frac{1+\delta}{2}V^2(\pi_{f_m},\pi_{\hat f_m})\mathbb{1}_{\Omega_m(\delta)}, \tag{6.12}$$

where $V^2(\pi_{f_m},\pi_{\hat f_m})$ is defined by

$$\begin{aligned}V^2(\pi_{f_m},\pi_{\hat f_m}) = {} & \frac{1}{n}\sum_{J\in m}|J|\frac{[\pi^{(J)}_{\hat f_m}-\pi^{(J)}_{f_m}]^2}{\pi^{(J)}_{f_m}}\left[\frac{\log[\pi^{(J)}_{\hat f_m}/\pi^{(J)}_{f_m}]}{\pi^{(J)}_{\hat f_m}/\pi^{(J)}_{f_m}-1}\right]^2 \\ & + \frac{1}{n}\sum_{J\in m}|J|\frac{[\pi^{(J)}_{\hat f_m}-\pi^{(J)}_{f_m}]^2}{1-\pi^{(J)}_{f_m}}\left[\frac{\log[(1-\pi^{(J)}_{\hat f_m})/(1-\pi^{(J)}_{f_m})]}{(1-\pi^{(J)}_{\hat f_m})/(1-\pi^{(J)}_{f_m})-1}\right]^2.\end{aligned} \tag{6.13}$$

Now we use that, for all $x>0$,

$$\frac{1}{1\vee x} \leqslant \frac{\log(x)}{x-1} \leqslant \frac{1}{1\wedge x}. \tag{6.14}$$

Hence we infer that

$$\frac{1}{(1+\delta)^2}\mathcal{X}^2_m\mathbb{1}_{\Omega_m(\delta)} \leqslant V^2(\pi_{f_m},\pi_{\hat f_m})\mathbb{1}_{\Omega_m(\delta)} \leqslant \frac{1}{(1-\delta)^2}\mathcal{X}^2_m\mathbb{1}_{\Omega_m(\delta)},$$

with $\mathcal{X}^2_m$ defined in (6.9). This entails that (6.8) is proved. It remains now to check that

$$\frac{4\rho^2|m|}{n} \leqslant \mathbb{E}_{f_0}[\mathcal{X}^2_m] \leqslant \frac{2|m|}{n}.$$

According to Lemma 4.1 , for all partition $J\in m$ and for any $x_i\in J$,

$$\pi_{\hat f_m}(x_i)=\pi^{(J)}_{\hat f_m}, \qquad \text{with} \qquad \pi^{(J)}_{\hat f_m}=\frac{1}{|J|}\sum_{i\in J}Y_i,$$

$$\text{and } \pi_{f_m}(x_i)=\pi^{(J)}_{f_m}, \qquad \text{with} \qquad \pi^{(J)}_{f_m}=\frac{1}{|J|}\sum_{i\in J}\pi_{f_0}(x_i).$$

Consequently,

$$\mathcal{X}_m^2 = \frac{1}{n}\sum_{J\in m}|J|\frac{(\sum_{k\in J}\varepsilon_k)^2}{\sum_{k\in J}\pi_{f_0}(x_k)[|J|-\sum_{k\in J}\pi_{f_0}(x_k)]} = \frac{1}{n}\sum_{J\in m}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]},$$

and finally

$$\mathbb{E}_{f_0}(\mathcal{X}_m^2) = \frac{1}{n}\sum_{J\in m}\mathbb{E}\left(\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]}\right) = \frac{1}{n}\sum_{J\in m}\left(\frac{1}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]}\right)\sum_{k\in J}\mathrm{Var}\,(Y_k)\,.$$

Consequently

$$\mathbb{E}_{f_0}(\mathcal{X}_m^2) = \frac{1}{n}\sum_{J\in m}\frac{\sum_{i\in J}\pi_{f_0}(x_i)(1-\pi_{f_0}(x_i))}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]}.$$

Now, according to Assumption $(\mathbf{A_2})$, and Lemma 4.1, for all partition m, all $J\in m$, and all $x_i\in J$

$$0<\rho^2\leqslant \pi_{f_0}(x_i)(1-\pi_{f_0}(x_i))\leqslant 1/4, \text{ and } 0<\rho\leqslant \pi_{f_m}^{(J)} \text{ and } 0<\rho\leqslant(1-\pi_{f_m}^{(J)}).$$

It follows that

$$4\rho^2 \leqslant \frac{\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]} = \frac{\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))}{|J|\pi_{f_m}^{(J)}} + \frac{\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))}{|J|[1-\pi_{f_m}^{(J)}]} \leqslant 2,$$

and thus

$$\frac{4\rho^2|m|}{n}\leqslant \frac{1}{n}\sum_{J\in m}\frac{\sum_{i\in J}\pi_{f_0}(x_i)(1-\pi_{f_0}(x_i))}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]}\leqslant\frac{2|m|}{n}.$$

In other words,

$$\frac{4\rho^2|m|}{n}\leqslant \mathbb{E}_{f_0}(\mathcal{X}_m^2)\leqslant\frac{2|m|}{n}.$$

The ends up the proof of (6.8) and (6.9).
• Proof of (6.10) : We start from (6.7), apply Assumption $(\mathbf{A_2})$ and Lemma 4.1, to obtain that and (6.10) is checked since

$$\begin{aligned}|\mathbb{E}\left(\mathcal{K}(\mathbb{P}_{f_m}^{(n)},\mathbb{P}_{\hat f_m}^{(n)})\mathbb{1}_{\Omega_m^c(\delta)}\right)| &\leqslant \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|\left[\log\left(\frac{\pi_{f_m}(x_i)}{\pi_{\hat f_m}(x_i)}\right)\mathbb{1}_{\Omega_m^c(\delta)}\right]\right| + \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|\left[\log\left(\frac{(1-\pi_{f_m}(x_i))}{(1-\pi_{\hat f_m}(x_i))}\right)\mathbb{1}_{\Omega_m^c(\delta)}\right]\right|\\ &\leqslant 2\log\left(\frac{1}{\rho}\right)\mathbb{P}[\Omega_m^c(\delta)].\end{aligned}$$

• Proof of (6.11): We come to the control of $\mathbb{P}_{f_0}[\Omega_m^c(\delta)]$. Since

$$\mathbb{P}[\Omega_m^c(\delta)] \leqslant \sum_{J\in m}\mathbb{P}\left\{\left|\frac{\pi^{(J)}_{\hat{f}_m}}{\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\}+\sum_{J\in m}\mathbb{P}\left\{\left|\frac{1-\pi^{(J)}_{\hat{f}_m}}{1-\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\},$$

by applying Lemma 4.1, we infer that

$$\mathbb{P}\left\{\left|\frac{\pi^{(J)}_{\hat{f}_m}}{\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\} = \mathbb{P}\left\{\left|\frac{\sum_{k\in J}\varepsilon_k}{\sum_{k\in J}\pi_{f_0}(x_k)}\right|\geqslant\delta\right\}=\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)\right\},$$

and

$$\mathbb{P}\left\{\left|\frac{1-\pi^{(J)}_{\hat{f}_m}}{1-\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\} = \mathbb{P}\left\{\left|\frac{\sum_{k\in J}\varepsilon_k}{\sum_{k\in J}(1-\pi_{f_0}(x_k))}\right|\geqslant\delta\right\}=\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}(1-\pi_{f_0}(x_k))\right\}.$$

We write

$$\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)\right\}\leqslant\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}$$

and

$$\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}(1-\pi_{f_0}(x_k))\right\}\leqslant\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}.$$

Then we have

$$\mathbb{P}[\Omega_m^c(\delta)] \leqslant 2\sum_{J\in m}\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}.$$

Now, we apply Bernstein Concentration Inequality (see Massart (2007) for example) to the right hand side of previous inequality, starting by recalling this Bernstein inequality.

Theorem 6.1. *Let $Z_1,\cdots,Z_n$ be independent real valued random variables. Assume that there exist some positive numbers v and c such that for all $k\geqslant 2$,*

$$\sum_{i=1}^n\mathbb{E}\left[|Z_i|^k\right]\leqslant\frac{k!}{2}vc^{k-2}.$$

Then for any positive z,

$$\mathbb{P}\left(\sum_{i=1}^n(Z_i-\mathbb{E}(Z_i)\geqslant\sqrt{2vz}+cz\right)\leqslant\exp(-z),\ \textit{and}\ \mathbb{P}\left(\sum_{i=1}^n(Z_i-\mathbb{E}(Z_i)\geqslant z\right)\leqslant\exp\left(-\frac{z^2}{2(v+cz)}\right).$$

Especially, if $|Z_i| \leqslant b$ *for all i, then*

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i) \geqslant z\right) \leqslant \exp\left(-\frac{z^2}{2(\sum_{i=1}^{n}\mathbb{E}(Z_i^2) + bz/3)}\right). \tag{6.15}$$

Applying (6.15) with $z = \delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))$, $b = 1$ and $v = \sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))$, we get that

$$\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right| \geqslant \delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}$$

is less than

$$2\exp\left(-\frac{\delta^2[\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))]^2}{2\left(\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k)) + (\delta/3)\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right)}\right),$$

and consequently

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right| \geqslant \delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\} &\leqslant 2\exp\left[-\frac{\delta^2}{2(1+\delta/3)}\left(\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right)\right]\\ &\leqslant 2\exp\left[-\frac{\delta^2}{2(1+\delta/3)}|J|\rho^2\right].\end{aligned}$$

Consequently,

$$\mathbb{P}[\Omega_m^c(\delta)] \quad \leqslant \quad 4|m|\exp(-\Delta\rho^2\Gamma[\log(n)]^2), \qquad \text{with} \qquad \Delta = \frac{\delta^2}{2(1+\delta/3)},$$

where Γ is given by (4.1). For $\epsilon > 0$ and δ such that

$$\frac{\delta^2}{2(1+\delta/3)}\rho^2\Gamma\log(n) \geqslant 2+\epsilon, \tag{6.16}$$

using that $|m| \leqslant n$ implies that

$$4|m|\exp\left(-\frac{\delta^2}{2(1+\delta/3)}\rho^2\Gamma[\log(n)]^2\right) \leqslant \frac{\kappa}{n^{(1+\epsilon)}}.$$

And Result (6.11) follows.

6.5. **Proof of Theorem 4.1.**

By definition, for all $m \in \mathcal{M}$,

$$\gamma_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leqslant \gamma_n(\hat{f}_m) + \text{pen}(m) \leqslant \gamma_n(f_m) + \text{pen}(m).$$

Applying Formula (6.1), we have

$$\gamma(\hat{f}_{\hat{m}}) - \gamma(f_0) \leqslant \gamma(f_m) - \gamma(f_0) + \langle\vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m\rangle_n + \text{pen}(m) - \text{pen}(\hat{m}). \tag{6.17}$$

Following Baraud (2000b) or Castellan (2003b), instead of bounding the supremum of the empirical process $\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n$, we split it in three terms. Let

$$\overline{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}_{f_0}(\gamma_n(t)) = - < \vec{\varepsilon}, f >_n$$

with $< \vec{\varepsilon}, f >_n$ defined in (6.1), and write

$$\begin{aligned}\gamma(\hat{f}_{\hat{m}}) - \gamma(f_0) &\leqslant \gamma(f_m) - \gamma(f_0) + \text{pen}(m) - \text{pen}(\hat{m}) \\ &\quad + \overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0) + \overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{\hat{m}}) + \overline{\gamma}_n(f_{\hat{m}}) - \overline{\gamma}_n(\hat{f}_{\hat{m}}).\end{aligned}$$

In other words,

$$\begin{aligned}\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) &\leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \text{pen}(m) - \text{pen}(\hat{m}) \\ &\quad + \overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0) + \overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{\hat{m}}) + \overline{\gamma}_n(f_{\hat{m}}) - \overline{\gamma}_n(\hat{f}_{\hat{m}}).\end{aligned} \tag{6.18}$$

The proof of Theorem 4.1 can be decomposed in three steps :

(R-1) We prove that for $\epsilon > 0$,

$$\mathbb{E}_{f_0}[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)}] \leqslant \frac{\kappa'(\rho, \delta, \Gamma, \epsilon)}{n^{(1+\epsilon)}}.$$

(R-2) Let $\Omega_1(\xi)$ be the event

$$\Omega_1(\xi) = \bigcap_{m' \in \mathcal{M}} \left\{ \chi^2_{m'} \mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant \frac{2}{n}|m'| + \frac{16}{n}\Big(1 + \frac{\delta}{3}\Big)\sqrt{(L_{m'}|m'| + \xi)|m'|} + \frac{8}{n}\Big(1 + \frac{\delta}{3}\Big)(L_{m'}|m'| + \xi) \right\},$$

where $(L_{m'})_{m' \in \mathcal{M}}$ satisfies Condition (4.2) and m_f is given by Definition 4.1. For all m' in $\mathcal{M}$ we prove that on $\Omega_1(\xi)$

$$\begin{aligned}\Big(\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'})\Big)\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant & \quad \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)|m'|\Big[2 + \Big(1 + \frac{\delta}{3}\Big)\Big(2\delta + 8L_{m'} + 16\sqrt{L_{m'}}\Big)\Big] \\ & + \frac{4\xi}{n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big(1 + \frac{\delta}{3}\Big)\Big(1 + \frac{4}{\delta}\Big) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)},\end{aligned} \tag{6.19}$$

and

$$\mathbb{P}(\Omega_1(\xi)^c) \leqslant 2\Sigma e^{-\xi}. \tag{6.20}$$

(R-3) Let $\Omega_2(\xi)$ be the event

$$\Omega_2(\xi) = \bigcap_{m' \in \mathcal{M}} \left[(\overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{m'})) \leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_{m'}}) - 2h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_{m'}}) + \frac{2}{n}(L'_m|m'| + \xi) \right].$$

We prove that, $\mathbb{P}(\Omega_2(\xi)^c) \leqslant \Sigma e^{-\xi}$.

Now, we will prove the result of Theorem 4.1 using (R-1), (R-2) and (R-3).
According to (6.18), we can write

$$\begin{aligned}\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \text{pen}(m) - \text{pen}(\hat{m}) \\ &\quad + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} + (\overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{\hat{m}}))\mathbb{1}_{\Omega_{m_f}(\delta)} + (\overline{\gamma}_n(f_{\hat{m}}) - \overline{\gamma}_n(\hat{f}_{\hat{m}})\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

Combining (R-2) and (R-3) with $m' = \hat{m}$, we infer that on $\Omega_1(\xi)\bigcap\Omega_2(\xi)$

$$
\begin{aligned}
\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant\ & \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\
& + \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)|\hat{m}|\Big[2 + \Big(1+\frac{\delta}{3}\Big)\Big(2\delta + 8L_{\hat{m}} + 16\sqrt{L_{\hat{m}}}\Big)\Big] + 2L_{\hat{m}}\frac{|\hat{m}|}{n} \\
& + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\
& + \Big[\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) - 2h^2(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{\hat{m}}},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant\ & \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\
& + \frac{|\hat{m}|}{n}\Big[\Big(\frac{1+\delta}{1-\delta}\Big) + \Big(\frac{\delta(1+\delta)^2}{1-\delta}\Big) + \Big(\frac{(1+\delta)^2}{1-\delta}\Big)\Big(6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big)\Big] \\
& + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\
& + \Big[\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) - 2h^2(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}})) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{\hat{m}}},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.
\end{aligned}
$$

Since

$$\Big\{\Big(\frac{1+\delta}{1-\delta}\Big)(1+\delta(1+\delta)) \vee \Big(\frac{(1+\delta)^2}{1-\delta}\Big)\Big\} \leqslant C(\delta) \text{ with } C(\delta) := \Big(\frac{1+\delta}{1-\delta}\Big)^3,$$

we infer

$$
\begin{aligned}
\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant\ & \mathcal{K}(\mathbb{P}^{(n)}_{f},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\
& + \frac{|\hat{m}|}{n}C(\delta)\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big] + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\
& + \Big[\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) - 2h^2(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{\hat{m}}},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.
\end{aligned}
$$

Using Pythagore's type identity $\mathcal{K}(\mathbb{P}_{f_0},\mathbb{P}_{\hat{f}_{\hat{m}}}) = \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) + \mathcal{K}(\mathbb{P}^{(n)}_{f_{\hat{m}}},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})$ (see Equation (7.42) in Massart (2007)) we have

$$
\begin{aligned}
\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant\ & \mathcal{K}(\mathbb{P}^{(n)}_{f},\mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\
& + \frac{|\hat{m}|}{n}C(\delta)\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big] + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\
& + \Big[\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) - 2h^2(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f_{\hat{m}}}) - \frac{\delta}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{\hat{m}}},\mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.
\end{aligned}
$$

Now, we successively use

(i) the relation between Kullback-Leibler information and the Hellinger distance $\mathcal{K}(\mathbb{P}^{(n)}_{f_{\hat{m}}}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \geq 2h^2(\mathbb{P}^{(n)}_{f_{\hat{m}}}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})$ (see Lemma 7.23 in Massart (2007)),

(ii) and inequality $h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \leqslant 2[h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_{\hat{m}}}) + h^2(\mathbb{P}^{(n)}_{f_{\hat{m}}}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})]$.

Consequently, on $\Omega_1(\xi) \bigcap \Omega_2(\xi)$

$$\begin{aligned}\frac{\delta}{1+\delta} h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}}) \mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0)) \mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\quad + \frac{|\hat{m}|}{n} C(\delta) \Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big] + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1 + \frac{\delta}{3}\Big)\Big(1 + \frac{4}{\delta}\Big)\Big].\end{aligned}$$

Since $\mathrm{pen}(\hat{m}) \geq \mu|\hat{m}|\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big]/n$, by taking $\mu = C(\delta)$ yields that on $\Omega_1(\xi) \bigcap \Omega_2(\xi)$

$$h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}}) \mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant \frac{2\mu^{1/3}}{\mu^{1/3} - 1}\Big(\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0)) \mathbb{1}_{\Omega_{m_f}(\delta)}\Big) + \frac{\xi}{n} C_1(\mu).$$

Then, using that

$$\mathbb{P}(\Omega_1(\xi)^c \cup \Omega_2(\xi)^c) \leqslant 3\Sigma e^{-\xi},$$

we deduce that $\mathbb{P}(\Omega_1(\xi) \cap \Omega_2(\xi)) \geq 1 - 3\Sigma e^{-\xi}$. We now integrating with respect to ξ, and use (R-1) to write that

$$\mathbb{E}_{f_0}\Big[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}}) \mathbb{1}_{\Omega_{m_f}(\delta)}\Big] \leqslant \frac{2\mu^{1/3}}{\mu^{1/3} - 1}\Big(\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\Big) + \frac{\kappa_1(\rho, \mu, \Gamma, \epsilon)}{n^{(1+\epsilon)}} + \frac{C_2(\mu, \Sigma)}{n}.$$

Furthermore, since $h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}}) \leqslant 1$, by applying Inequality (6.11) we have,

$$\mathbb{E}_{f_0}\Big[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}}) \mathbb{1}_{\Omega^c_{m_f}(\delta)}\Big] \leq \frac{\kappa_2(\rho, \mu, \Gamma, \epsilon)}{n^{(1+\epsilon)}}.$$

Hence we conclude that

$$\mathbb{E}_{f_0}\Big[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}})\Big] \leqslant \frac{2\mu^{1/3}}{\mu^{1/3} - 1}\Big(\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\Big) + \frac{\kappa_3(\rho, \mu, \Gamma, \epsilon)}{n^{(1+\epsilon)}} + \frac{C_2(\mu, \Sigma)}{n},$$

and minimizing over $\mathcal{M}$ leads to the result of Theorem 4.1.
We now come to the proofs of (R-1), (R-2) and (R-3).
• Proof of (R-1)
We know that

$$\begin{aligned}\Big|\mathbb{E}_{f_0}\Big[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0)) \mathbb{1}_{\Omega_{m_f}(\delta)}\Big]\Big| &= \Big|\mathbb{E}_{f_0}\Big[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0)) \mathbb{1}_{\Omega^c_{m_f}(\delta)}\Big]\Big| \\ &\leq \mathbb{E}_{f_0}\Big[\frac{1}{n}\sum_{i=1}^n \Big\{\Big|\epsilon_i \log\{\frac{\pi_{f_m}(x_i)}{\pi_{f_0}(x_i)}\}\Big| + \Big|\epsilon_i \log\{\frac{1 - \pi_{f_m}(x_i)}{1 - \pi_{f_0}(x_i)}\}\Big|\Big\} \mathbb{1}_{\Omega^c_{m_f}(\delta)}\Big] \\ &\leq 2\log\Big\{\frac{1}{\rho}\Big\} \mathbb{P}(\Omega^c_{m_f}(\delta)).\end{aligned}$$

We conclude the proof of (R-1) by using Inequality (6.11), which implies that

$$\Big|\mathbb{E}_{f_0}\Big[(\overline{\gamma}_n(f_m)-\overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)}\Big]\Big| \leq 2\log\left\{\frac{1}{\rho}\right\}\frac{\kappa(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}} = \frac{\kappa'(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}}.$$

- Proof of (R-2)

We start by the proof of (6.19)

$$\begin{aligned}\overline{\gamma}_n(f_{m'})-\overline{\gamma}_n(\hat{f}_{m'}) &= -\frac{1}{n}\sum_{i=1}^{n}\Big\{\epsilon_i\log\Big(\frac{\pi_{f_{m'}}(x_i)}{\pi_{\hat{f}_{m'}}(x_i)}\Big)-\epsilon_i\log\Big(\frac{1-\pi_{f_{m'}}(x_i)}{1-\pi_{\hat{f}_{m'}}(x_i)}\Big)\Big\}\\ &= -\frac{1}{n}\sum_{J\in m'}\Big(\sum_{i\in J}\epsilon_i\Big)\Big[\frac{\sqrt{|J|\pi^{(J)}_{f_{m'}}}}{\sqrt{|J|\pi^{(J)}_{f_{m'}}}}\log\Big(\frac{\pi^{(J)}_{f_{m'}}}{\pi^{(J)}_{\hat{f}_{m'}}}\Big)-\frac{\sqrt{|J|1-\pi^{(J)}_{f_{m'}}}}{\sqrt{|J|(1-\pi^{(J)}_{f_{m'}})}}\log\Big(\frac{1-\pi^{(J)}_{f_{m'}}}{1-\pi^{(J)}_{\hat{f}_{m'}}}\Big)\Big].\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}\overline{\gamma}_n(f_{m'})-\overline{\gamma}_n(\hat{f}_{m'}) \leq &\sqrt{\frac{1}{n}\sum_{J\in m'}|J|\Big[\pi^{(J)}_{f_{m'}}\log^2\Big(\frac{\pi^{(J)}_{\hat{f}_{m'}}}{\pi^{(J)}_{f_{m'}}}\Big)+(1-\pi^{(J)}_{f_{m'}})\log^2\Big(\frac{1-\pi^{(J)}_{\hat{f}_{m'}}}{1-\pi^{(J)}_{f_{m'}}}\Big)\Big]}\\ &\times\sqrt{\frac{1}{n}\sum_{J\in m'}\Big[\frac{\Big(\sum_{i\in J}\epsilon_i\Big)^2}{|J|\pi^{(J)}_{f_{m'}}}+\frac{\Big(\sum_{i\in J}\epsilon_i\Big)^2}{|J|(1-\pi^{(J)}_{f_{m'}})}\Big]},\end{aligned}$$

and in other words

$$\overline{\gamma}_n(f_{m'})-\overline{\gamma}_n(\hat{f}_{m'}) \leq \sqrt{\mathcal{X}^2_{m'}}\times\sqrt{V^2(\pi_{f_{m'}},\pi_{\hat{f}_{m'}})},$$

where $\mathcal{X}^2_{m'}$ and $V^2(\pi_{f_{m'}},\pi_{\hat{f}_{m'}})$ are defined respectively in (6.9) and (6.13) . Using both that inequality $2xy\leqslant\theta x^2+\theta^{-1}y^2$, for all $x>0$, $y>0$ with $\theta=(1+\delta)/(1-\delta)$, and Inequality (6.12), we obtain on $\Omega_{m_f}(\delta)$ that,

$$\overline{\gamma}_n(f_{m'})-\overline{\gamma}_n(\hat{f}_{m'})) \leq \frac{1}{2}\Big(\frac{1+\delta}{1-\delta}\Big)\chi^2_{m'}+\frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}},\mathbb{P}^{(n)}_{\hat{f}_{m'}}).$$

Consequently, on $\Omega_1(\xi)$

$$\begin{aligned}(\overline{\gamma}_n(f_{m'})-\overline{\gamma}_n(\hat{f}_{m'}))\mathbb{1}_{\Omega_{m_f}(\delta)} &\leq \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big[2|m'|+16\Big(1+\frac{\delta}{3}\Big)\sqrt{(L_{m'}|m'|+\xi)|m'|}+8\Big(1+\frac{\delta}{3}\Big)(L_{m'}|m'|+\xi)\Big]\\ &+ \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}},\mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

Using inequalities $|x+y|^{1/2} \leqslant |x|^{1/2} + |y|^{1/2}$ and $2xy \leqslant \theta x^2 + \theta^{-1} y^2$ with $\theta = \delta/4$, we infer that (6.19) follows since

$$\begin{aligned}
\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'}))\mathbb{1}_{\Omega_{m_f}(\delta)} &\le \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big[2|m'| + \Big(1+\frac{\delta}{3}\Big)\Big(16\sqrt{L_{m'}}|m'| + 8L_{m'}|m'| + 2\delta|m'|\Big) \\
&\quad +8\xi\Big(1+\frac{\delta}{3}\Big)(1+\frac{4}{\delta})\Big] + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)} \\
&\le \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)|m'|\Big[2 + \Big(1+\frac{\delta}{3}\Big)\Big(2\delta + 8L_{m'} + 16\sqrt{L_{m'}}\Big)\Big] \\
&\quad + \frac{4\xi}{n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)}.
\end{aligned}$$

• Proof of (6.20) :
Write $\mathcal{X}^2_{m'} = \sum_{J\in m'}\{Z_{1,J} + Z_{2,J}\}$, where

$$Z_{1,J} = \frac{1}{n}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi^{(J)}_{f_{m'}}} \text{ and } Z_{2,J} = \frac{1}{n}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|(1-\pi^{(J)}_{f_{m'}})}.$$

We will control $\sum_{J\in m'} Z_{1,J}$ and $\sum_{J\in m'} Z_{2,J}$ separately. In order to use Bernstein inequality (see Theorem 6.1), we need an upper bound of $\sum_{J\in m'}\mathbb{E}[Z^p_{1,J}\mathbb{1}_{\Omega_{m_f}(\delta)}]$, for every $p \ge 2$. By definition

$$\mathbb{E}[Z^p_{1,J}\mathbb{1}_{\Omega_{m_f}(\delta)}] = \frac{1}{\Big(n|J|\pi^{(J)}_{f_{m'}}\Big)^p}\int_0^\infty 2px^{2p-1}\mathbb{P}\Big(\Big\{|\sum_{k\in J}\varepsilon_k| \ge x\Big\}\cap\Omega_{m_f}(\delta)\Big)dx.$$

For every m' constructed on the grid m_f, for all $J\in m'$, on $\Omega_{m_f}(\delta)\cap\Big\{x \leqslant |\sum_{k\in J}\varepsilon_k|\Big\}$, we have

$$x \leqslant |\sum_{k\in J}\varepsilon_k| \leqslant \delta\sum_{i\in J}\pi_{f_0}(x_i).$$

Combining the previous inequality, the Bernstein inequality (6.15) with the fact that $\varepsilon_k \leqslant 1$, we infer that

$$\begin{aligned}
\mathbb{E}[Z^p_{1,J}\mathbb{1}_{\Omega_{m_f}(\delta)}] &\le \frac{1}{\Big(n\sum_{k\in J}\pi_{f_0}(x_k)\Big)^p}\int_0^{\delta\sum_{k\in J}\pi_{f_0}(x_k)} 2px^{2p-1}\mathbb{P}\Big(|\sum_{k\in J}\varepsilon_k| \ge x\Big)dx \\
&\le \frac{1}{\Big(n\sum_{k\in J}\pi_{f_0}(x_k)\Big)^p}\int_0^{\delta\sum_{i\in J}\pi_{f_0}(x_i)} 4px^{2p-1}\exp\Big(-\frac{x^2}{2\Big(\frac{x}{3}+\sum_{k\in J}\pi_{f_0}(x_k)\Big)}\Big)dx \\
&\le \frac{1}{\Big(n\sum_{k\in J}\pi_{f_0}(x_k)\Big)^p}\int_0^{\delta\sum_{i\in J}\pi_{f_0}(x_i)} 4px^{2p-1}\exp\Big(-\frac{x^2}{2\Big(1+\frac{\delta}{3}\Big)\sum_{k\in J}\pi_{f_0}(x_k)}\Big)dx \\
&\le \frac{1}{n^p}2^{p+1}(1+\frac{\delta}{3})^p p\int_0^\infty t^{p-1}\exp(-t)dt \\
&\le \frac{1}{n^p}2^{p+1}p(1+\frac{\delta}{3})^p(p!).
\end{aligned}$$

Consequently

$$\sum_{J\in m'} \mathbb{E}[Z_{1,J}^p \mathbb{I}_{\Omega_{m_f}(\delta)}] \leqslant \frac{1}{n^p} 2^{p+1} p(1+\frac{\delta}{3})^p (p!) \times |m'|.$$

Now, since $p \leqslant 2^{p-1}$, we have

$$\sum_{J\in m'} \mathbb{E}[Z_{1,J}^p \mathbb{I}_{\Omega_{m_f}(\delta)}] \leqslant \frac{p!}{2} \times \Big[\frac{32}{n^2}(1+\frac{\delta}{3})^2 |m'|\Big] \times \Big[\frac{4}{n}(1+\frac{\delta}{3})\Big]^{p-2}.$$

Using Bernstein inequality and that $\mathbb{E}\Big[\sum_{J\in m'} Z_{1,J})\Big] \leqslant |m'|/n$, we have that for every positive x

$$\mathbb{P}\Big(\sum_{J\in m'} Z_{1,J} \mathbb{I}_{\Omega_{m_f}(\delta)} \geq \frac{|m'|}{n} + \frac{8}{n}(1+\frac{\delta}{3})\sqrt{x|m'|} + \frac{4}{n}(1+\frac{\delta}{3})x\Big) \leqslant \exp(-x).$$

In the same way we prove that

$$\mathbb{P}\Big(\sum_{J\in m'} Z_{2,J} \mathbb{I}_{\Omega_{m_f}(\delta)} \geq \frac{|m'|}{n} + \frac{8}{n}(1+\frac{\delta}{3})\sqrt{x|m'|} + \frac{4}{n}(1+\frac{\delta}{3})x\Big) \leqslant \exp(-x).$$

Hence

$$\mathbb{P}\Big(\mathcal{X}_{m'}^2 \mathbb{I}_{\Omega_{m_f}(\delta)} \geq \frac{2|m'|}{n} + \frac{16}{n}(1+\frac{\delta}{3})\sqrt{x|m'|} + \frac{8}{n}(1+\frac{\delta}{3})x\Big) \leqslant 2\exp(-x),$$

and we conclude that $\mathbb{P}(\Omega_1^c(\xi)) \leqslant 2\sum_{m'} \exp(-L'_m|m'| - \xi) = 2\Sigma e^{-\xi}$. This ends the proof of (R-2).

• Proof of (R-3)

Recall that $\overline{\gamma}_n(f) = \gamma_n(f) - \mathbb{E}(\gamma_n(f))$ for every f. According to Markov inequality, for $b > 0$,

$$\begin{aligned}
\mathbb{P}((\overline{\gamma}_n(f_0) - \overline{\gamma}_n(g)) \geq b) &= \mathbb{P}\Big(\exp\Big(\frac{n}{2}(\overline{\gamma}_n(f_0) - \overline{\gamma}_n(g))\Big) \geq \exp\Big(\frac{nb}{2}\Big)\Big) \\
&\leq \exp\Big(\frac{-nb}{2}\Big)\mathbb{E}\Big[\exp\Big(\frac{n}{2}(\overline{\gamma}_n(f_0) - \overline{\gamma}_n(g))\Big)\Big] \\
&= \exp\Big[\frac{-nb}{2} + \log\mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0) - \gamma_n(g)\Big) + \frac{n}{2}\mathbb{E}\Big[\gamma_n(g) - \gamma_n(f_0)\Big]\Big)\Big] \\
&\leq \exp\Big[\frac{-nb}{2} + \frac{n}{2}\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)}) + \log\mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0) - \gamma_n(g)\Big)\Big)\Big]\Big].
\end{aligned}$$

Now,

$$\begin{aligned}
\log \mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big)\Big] &= \log \mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{i=1}^{n} Y_i \log(\frac{\pi_g(x_i)}{\pi_{f_0}(x_i)}) + (1-Y_i)\log(\frac{1-\pi_g(x_i)}{1-\pi_{f_0}(x_i)})\Big)\Big] \\
&= \log \mathbb{E}\Big[\Pi_{i=1}^n\Big\{\Big(\frac{\pi_g(x_i)}{\pi_{f_0}(x_i)}\Big)^{Y_i/2}\times\Big(\frac{1-\pi_g(x_i)}{1-\pi_{f_0}(x_i)}\Big)^{(1-Y_i)/2}\Big\}\Big] \\
&= \log \Pi_{i=1}^n\Big\{\sqrt{\frac{\pi_g(x_i)}{\pi_{f_0}(x_i)}}\pi_{f_0}(x_i) + \sqrt{\frac{1-\pi_g(x_i)}{1-\pi_{f_0}(x_i)}}(1-\pi_{f_0}(x_i))\Big\} \\
&= \sum_{i=1}^{n}\log\Big\{\sqrt{\pi_g(x_i)\pi_{f_0}(x_i)} + \sqrt{(1-\pi_g(x_i))(1-\pi_{f_0}(x_i))}\Big\}.
\end{aligned}$$

In other words we have

$$\log \mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big) = \sum_{i=1}^{n}\log\Big\{1-\frac{1}{2}\Big[\Big(\sqrt{\pi_{f_0}(x_i)}-\sqrt{\pi_g(x_i)}\Big)^2+\Big(\sqrt{1-\pi_{f_0}(x_i)}-\sqrt{1-\pi_g(x_i)}\Big)^2\Big]\Big\}.$$

This implies that

$$\begin{aligned}
\log \mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big)\Big] &\leq \sum_{i=1}^{n} -\frac{1}{2}\Big[\Big(\sqrt{\pi_{f_0}(x_i)}-\sqrt{\pi_g(x_i)}\Big)^2+\Big(\sqrt{1-\pi_{f_0}(x_i)}-\sqrt{1-\pi_g(x_i)}\Big)^2\Big] \\
&= -nh^2(\mathbb{P}_{f_0},\mathbb{P}_g).
\end{aligned}$$

Consequently

$$\mathbb{P}(\overline{\gamma}_n(f_0)-\overline{\gamma}_n(g)\geq b)\leqslant \exp\Big[\frac{-nb}{2}+\frac{n}{2}\mathcal{K}(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})-nh^2(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})\Big],$$

and, if we choose for positive x,

$$b=\frac{2x}{n}+\mathcal{K}(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})-2h^2(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})>0,$$

we have,

$$\mathbb{P}\Big(\overline{\gamma}_n(f_0)-\overline{\gamma}_n(g)\geq\frac{2x}{n}+\mathcal{K}(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})-2h^2(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})\Big)\leqslant\exp(-x).$$

We conclude that $\mathbb{P}(\Omega_2^c(\xi))\leqslant\sum_{m'}\exp(-L'_m|m'|-\xi)\leq\Sigma e^{-\xi}$, which ends the proof of (R-3). □

7. Appendix

7.1. **Proof of Lemma 4.1.** By definition

$$f_m=\arg\min_{f\in S_m}\Bigg[\sum_{i=1}^{n}\log(1+\exp(f(x_i)))-\pi_{f_0}(x_i)f(x_i)\Bigg].$$

For all $f \in S_m$, for all $J \in m$ and for all $x \in J$, we have $f(x) = f^{(J)}$. Hence $f_m(x) = \overline{f}_m^{(J)}$ for all x in J, and for all J in m, we aim at finding $\overline{f}_m^{(J)}$ such that

$$\overline{f}_m^{(J)} = \arg\min_{f^{(J)}} \left[|J| \log(1 + \exp(f^{(J)})) - \sum_{i \in J} \pi_{f_0}(x_i) f^{(J)} \right]$$

where $|J| = card\{i \in \{1, ..., n\}; x_i \in J\}$. Easy calculations show that he coefficient $\overline{f}_m^{(J)}$ satisfies

$$|J| \frac{\exp(\overline{f}_m^{(J)})}{1 + \exp(\overline{f}_m^{(J)})} - \sum_{i \in J} \pi_{f_0}(x_i) = 0,$$

that is

$$\overline{f}_m^{(J)} = \log\left(\frac{\sum_{i \in J} \pi_{f_0}(x_i)}{|J|(1 - \sum_{i \in J} \pi_{f_0}(x_i)/|J|)} \right). \tag{7.1}$$

Consequently, π_{f_m} defined as in (2.2) satisfies that $\pi_{f_m}(x) = \pi_{f_m}^{(J)}$ for all $x \in J$, where

$$\pi_{f_m}^{(J)} = \frac{1}{|J|} \sum_{i \in J} \pi_{f_0}(x_i),$$

and hence $\pi_{f_m} = \arg\min_{t \in S_m} \| t - \pi_{f_0} \|_n$ is the usual projection of π_{f_0} on to $S_m =< \Phi_j, j \in m >$. In the same way, $\hat{f}_m$ defined by (4.10) satisfies $\hat{f}_m(t) = \hat{f}_m^{(J)}$ for all $t \in J$, where

$$\hat{f}_m^{(J)} = \log\left(\frac{\sum_{i \in J} Y_i}{|J|(1 - \sum_{i \in J} Y_i/|J|)} \right).$$

In other words, $\pi_{\hat{f}_m}$, defined as π_f with f replaced by $\pi_{\hat{f}_m}$, satisfies $\pi_{\hat{f}_m}(x) = \pi_{\hat{f}_m}^{(J)}$, for all $x \in J$, with

$$\pi_{\hat{f}_m}^{(J)} = \frac{1}{|J|} \sum_{i \in J} Y_i.$$

7.2. **Proof of Lemma 6.2.** In the following, for the sake of notation simplicity, we will use $\gamma(\beta)$ for $\gamma(f_\beta)$. A second-order Taylor expansion of the function $\gamma()$ around β^* gives for any $\beta \in \Lambda_m$

$$\begin{aligned}\gamma(\beta) &= \gamma(\beta^*) + \nabla_\beta \gamma(\beta^*)(\beta - \beta^*) \\ &\quad + \int_0^1 (1-t) \sum_{i_1 + \dots + i_D = 2} \frac{2!}{i_1! \dots i_D!} (\beta_1 - \beta_1^*)^{i_1} \dots (\beta_D - \beta_D^*)^{i_D} \frac{\partial \gamma^2}{\partial \beta_1 \dots \partial \beta_D} (\beta^* + t(\beta - \beta *)) dt.\end{aligned}$$

Easy calculation shows that

$$
\begin{aligned}
& \sum_{i_1+\cdots+i_D=2} \frac{2!}{i_1!\dots i_D!}(\beta_1-\beta_1^*)^{i_1}\dots(\beta_D-\beta_D^*)^{i_D}\frac{\partial\gamma^2}{\partial\beta_1\dots\partial\beta_D}(\beta^*+t(\beta-\beta*)) \\
= & \sum_{j=1}^{D}\frac{1}{n}\sum_{i=1}^{n}\psi_j^2(x_i)(\beta_j-\beta_j^*)^2\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\left[1-\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\right] \\
+ & \ 2\sum_{l\neq k}\frac{1}{n}\sum_{i=1}^{n}\psi_l(x_i)\psi_k(x_i)(\beta_l-\beta_l^*)(\beta_k-\beta_k^*)\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\left[1-\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\right] \\
= & \ \frac{1}{n}\sum_{i=1}^{n}\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\left[1-\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\right](f_\beta(x_i)-f_{\beta^*}(x_i))^2.
\end{aligned}
$$

This implies that

$$\gamma(\beta)\geq\gamma(\beta^*)+\nabla_\beta\gamma(\beta^*)(\beta-\beta^*)+\frac{\mathcal{U}_0^2}{2}\|f_\beta-f_{\beta^*}\|_n^2.$$

Since β^* is the minimizer of $\gamma(.)$ over the set Λ_m, we have $\nabla_\beta\gamma(\beta^*)(\beta-\beta^*)\geq 0$ for all $\beta\in\Lambda_m$. Thus the result follows.

7.3. **Proof of Lemma 6.3.** Let S_D and $S_{D'}$ two vector spaces of dimension D and D' respectively. Set $S=S_D\cap\mathbb{L}_\infty(C_0)+S_{D'}\cap\mathbb{L}_\infty(C_0)$ and $\vec{\varepsilon}'$ be an independent copie of $\vec{\varepsilon}$. Set

$$Z=\sup_{u\in S}\frac{\langle\vec{\varepsilon},u\rangle_n}{\|u\|_n},\ \text{and for all } i=1,\dots,n,\quad Z^{(i)}=\sup_{u\in S}\frac{1}{\|u\|_n}\left(\frac{1}{n}\sum_{k\neq i}\varepsilon_k u(x_k)+\varepsilon_i' u(x_i)\right). \tag{7.2}$$

By Cauchy-Schwarz Inequality the supremum in (7.2) is achieved at $\Pi_S(\vec{\varepsilon})$. Consequently,

$$Z-Z^{(i)}\leqslant\frac{(\varepsilon_i-\varepsilon_i')(\Pi_S(\vec{\varepsilon})(x_i)}{n\|\Pi_S(\vec{\varepsilon})\|_n},\qquad\text{and}\qquad \mathbb{E}_{f_0}[(Z-Z^{(i)})^2|\vec{\varepsilon}]\leq\mathbb{E}_{f_0}\left[\frac{(\varepsilon_i-\varepsilon_i')^2[\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2\|\Pi_S(\vec{\varepsilon})\|_n^2}|\vec{\varepsilon}\right]$$

with

$$
\begin{aligned}
\mathbb{E}_{f_0}\left[\frac{(\varepsilon_i-\varepsilon_i')^2[\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2\|\Pi_S(\vec{\varepsilon})\|_n^2}|\vec{\varepsilon}\right] &= \frac{[\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2\|\Pi_S(\vec{\varepsilon})\|_n^2}\mathbb{E}_{f_0}\left[(\varepsilon_i-\varepsilon_i')^2|\vec{\varepsilon}\right] \\
&= \frac{[\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2\|\Pi_S(\vec{\varepsilon})\|_n^2}\left(\varepsilon_i^2+\mathbb{E}_{f_0}(\varepsilon_i^2)\right)\leq\frac{5[\Pi_S(\vec{\varepsilon})(x_i)]^2}{4n^2\|\Pi_S(\vec{\varepsilon})\|_n^2}.
\end{aligned}
$$

This implies that

$$\sum_{i=1}^{n}\mathbb{E}_{f_0}[(Z-Z^{(i)})^2\mathbb{1}_{Z>Z^{(i)}}|\vec{\varepsilon}]\leq\frac{5}{4n}.$$

We now apply Lemma 7.1 from Boucheron *et al.* (2004)), that is recalled here.

Lemma 7.1. *Let $X_1, \ldots, X_n$ independent random variables taking values in a measurable space $\mathcal{X}$. Denote by X_1^n the vector of these n random variables. Set $Z = f(X_1, \ldots, X_n)$ and $Z^{(i)} = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$, where $X_1', \ldots, X_n'$ denote independent copies of $X_1, \ldots, X_n$ and $f : \mathcal{X}^n \to \mathbb{R}$ some measurable function. Assume that there exists a positive constant c such that, $\mathbb{E}_{f_0}\left[\sum_{i=1}^n (Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} | X_1^n\right] \leqslant c$. Then for all $t > 0$,*

$$\mathbb{P}_{f_0}(Z > \mathbb{E}_{f_0}(Z) + t) \leqslant e^{-t^2/4c}.$$

Applying Lemma 7.1 to Z defined in (7.2), we obtain that for all $x > 0$,

$$\mathbb{P}\left(\sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} > \mathbb{E}_{f_0}\left[\sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n}\right] + \sqrt{\frac{5x}{n}}\right) \leqslant \exp(-x).$$

Let $\{\psi_1, \ldots, \psi_{D+D'}\}$ be an orthonormal basis of $S_D + S_{D'}$. Using Jensen's Inequality, we write

$$\begin{aligned}
\mathbb{E}_{f_0}\left[\sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n}\right] = \mathbb{E}_{f_0}(\| \Pi_S(\vec{\varepsilon}) \|_n) &= \mathbb{E}_{f_0}\left[\left(\sum_{k=1}^{D+D'} (\langle \vec{\varepsilon}, \psi_k \rangle_n)^2\right)^{1/2}\right] \\
&\leq \left(\sum_{k=1}^{D+D'} \mathbb{E}_{f_0}(\langle \vec{\varepsilon}, \psi_k \rangle_n)^2\right)^{1/2} \\
&\leqslant \sqrt{\frac{D + D'}{4n}}.
\end{aligned}$$

This concludes the proof of Lemma 6.3.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pp. 267–281. Akademinai Kiado.

Arlot, S. and P. Massart (2009). Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research 10*, 245–279.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics 4*, 384–414.

Baraud, Y. (2000a). Model selection for regression on a fixed design. *Probability Theory and Related Fields 117*(4), 467–493.

Baraud, Y. (2000b). Model selection for regression on a fixed design. *Probab. Theory Related Fields 117*(4), 467–493.

Baudry, J.-P., C. Maugis, and B. Michel (2012). Slope heuristics: overview and implementation. *Statistics and Computing 22*(2), 455–470.

Birgé, L. (2014). Model selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields 158*(3-4), 533–574.

Birgé, L. and P. Massart (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli 4*(3), 329–375.

Birgé, L. and P. Massart (2001). Gaussian model selection. *Journal of the European Mathematical Society 3*(3), 203–268.

Birgé, L. and P. Massart (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields 138*(1-2), 33–73.

Bontemps, D. and W. Toussile (2013). Clustering and variable selection for categorical multivariate data. *Electronic Journal of Statistics 7*, 2344–2371.

Boucheron, S., G. Lugosi, and O. Bousquet (2004). Concentration inequalities in machine learning summer school 2003. *Advanced Lectures on Machine Learning 3176*, 169–240.

Braun, J. V., R. Braun, and H.-G. Müller (2000). Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika 87*(2), 301–314.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via 1 and 1+ 2 penalization. *Electronic Journal of Statistics 2*, 1153–1194.

Castellan, G. (2003a). Density estimation via exponential model selection. *Information Theory, IEEE Transactions on 49*(8), 2052–2060.

Castellan, G. (2003b). Density estimation via exponential model selection. *IEEE Trans. Inform. Theory 49*(8), 2052–2060.

Cox, D. D. and F. O'Sullivan (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist. 18*(4), 1676–1695.

Durot, C., E. Lebarbier, and A.-S. Tocquet (2009). Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli 15*(2), 475–507.

Fan, J., M. Farmen, and I. Gijbels (1998). Local maximum likelihood estimation and inference. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*(3), 591–608.

Farmen, M. W. (1996). *The smoothed bootstrap for variable bandwidth selection and some results in nonparametric logistic regression*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–The University of North Carolina at Chapel Hill.

Hastie, T. J. (1983). NONPARAMETRIC LOGISTIC REGRESSION. *Appl. Stat.*.

Kwemou, M. (2012). Non-asymptotic oracle inequalities for the lasso and group lasso in high dimensional logistic model. Technical report, preprint arXiv:1206.0710.

Lebarbier, É. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing 85*(4), 717–736.

Lerasle, M. (2012). Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 48*(3), 884–908.

Lu, F. (2006). *Regularized nonparametric logistic regression and kernel regularization*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–The University of Wisconsin - Madison.

Massart, P. (2007). *Concentration inequalities and model selection*, Volume 1896 of *Lecture Notes in Mathematics*. Berlin: Springer. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Maugis, C. and B. Michel (2011). Data-driven penalty calibration: a case study for gaussian mixture model selection. *ESAIM: Probability and Statistics 15*, 320–339.

Raghavan, N. (1993). *Bayesian inference in nonparametric logistic regression*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of Illinois at Urbana-Champaign.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics 36*(2), 614–645.

Vexler, A. and G. Gurevich (2006). Guaranteed local maximum likelihood detection of a change point in nonparametric logistic regression. *Comm. Statist. Theory Methods 35*(4-6), 711–726.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica 9*(2), 475–499.

(1) Laboratoire de Mathématiques et de Modélisation d'Evry, Université d'Évry Val d'Essonne, UMR CNRS 8071- USC INRA, 23 Boulevard de France, 91037 Évry

(2) INRA, UR 341 MIA-Jouy,, Domaine de Vilvert,, F78352 Jouy-en-Josas, France